ADDRESSING THE GENDER GAP IN MATHEMATICS: THE IMPACT OF SUPPLEMENTARY SINGLE-SEX MATHEMATICS CLASSES ON MIDDLE SCHOOL STUDENTS' MATHEMATICS SELF-EFFICACY, SENSE OF BELONGING, AND ACHIEVEMENT

By

Susanna L. Brock

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Education

Baltimore, Maryland
October 2017

Abstract

The purpose of this mixed-methods study was to assess the gender gap in mathematics achievement, self-efficacy, and sense of belonging in a co-educational middle school context and to determine if an all-girls supplementary mathematics class, or "math workshop," could help to close this mathematics gender gap for high-ability students in fifth and seventh grades. A pre-test-post-test design with multiple comparison groups was used to assess the intervention using standardized test data and student survey responses. The results suggested that the intervention helped to close the gender gap between high-ability boys and girls who were enrolled in accelerated mathematics classes. High-ability girls gained more scaled score point on the annual standardized mathematics test than high-ability boys, $t(57)=-2.36$, $p=.022$, $d=.60$. Analysis of survey responses indicated that students who participated in the math workshop intervention saw a greater increase in sense of belonging scores compared to student who did not receive the intervention, $t(200)=2.299$, $p=.023$. However, this result was not significant by gender. Thus, the intervention appeared to have been mediated more by a change in students' sense of belonging in mathematics than their self-efficacy. Interviews with girls who participated in the intervention suggested that the opportunity to work with peers was a central component of their increased sense of belonging in this academic context.

Key Words: mathematics, achievement, gender gap, self-efficacy, sense of belonging

Dedication

I dedicate this dissertation to my parents, Mary Jane and Charles Brock. Dad: thank you for all the mornings together at the breakfast table reading the Science Times. You've given me the courage to ask tough questions and the grit to try and answer them. Mom: you named me after a great female school leader and have shown me through your actions the true wisdom of St. Catherine's motto: *What we keep we lose; only what we give remains our own*. This is for you.

Acknowledgements

There are many people without whom I could not have completed this project. I'd like to thank my advisor for my first two years, Christopher Sessums, for being an invaluable support as I began this journey. He also helped to connect me with my subsequent advisor Juliana Paré-Blagoev, who generously agreed to take on advising this research and has proven to be a dedicated and insightful mentor and guide through this process. I am also deeply grateful for the feedback and support of my committee members Karen Karp and Charol Shakeshaft. Their detailed readings of my work have provided me with critical feedback for improving my writing and communicating my ideas.

I am grateful for the funding provided for this research from the nonprofit organization FHI 360. This organization has been central in creating meaningful dialogue between researchers and practitioners regarding girls' mathematics identity, and I have benefited from their deep commitment to improving girls' experience in mathematics.

I would like to thank my colleagues at the BC School. I am deeply indebted for the incredible openness, curiosity, and collaboration that I have experienced each day working at this institution. In particular, I am very appreciative of the leadership team of Bob Vitalo and Jim Shapiro that supported this project and allowed me to design an intervention to address the gender gap. On a daily basis, I relied on the never-ending patience and assistance of the mathematics department team. Most essentially, my friend and colleague Lisa Hartmann who believed in this project and provided countless moments of encouragement and practical support. I also was lucky to work with an all-star team of math teachers who worked with me to make this idea a reality and to improve it along the way: Tom Jameson, Greg Scordato, Richard Blatherwick, Kathy Harrington, Kathy Grimes-Lamb, Shahna-Lee James, and Yabome Kabia.

Thank you most sincerely for all of the many ways large and small you each helped me achieve this finished product. Generous colleagues outside of my department also assisted me in this research. Elizabeth Hayward and Brandie Melendez provided crucial expertise and Nancy Holodak offered countless numbers of hours both discussing my methods and conducting interviews. Aidan Lucey and Nick Marchese advised me on the survey design and technology needs of the research. I also wish to acknowledge my colleagues and office mates Deborah Hickox and Elizabeth Epstein for their unwavering encouragement these past three years.

I'm also lucky to have friends who kept on cheering for me throughout this experience. Thank you Kate Harper, Hana Alberts, and Elaine Kelly. I owe a special nod to my friend Frank Provenzano for his help in advising me on statistical analyses. Within the Johns Hopkins program, thanks to my friends in the 2014 cohort, especially Grace Chu and Peling Li.

Thank you Mom, Dad and big brother Walker for raising me to believe that I could become "Dr. Brock" one day. When I doubted myself, I always knew the three of you would be there to set me straight.

Finally, I offer the greatest thank you to my husband and best friend, William Deringer. Thank you for the compromises you made these last three years so that I could pursue this dream. Your unquestioning confidence and support for my work made difficult moments manageable and exciting moments more joyful. Thank you, Willy; you are my true partner for life.

Table of Contents

List of Tables

List of Figures

CHAPTER 1. EXECUTIVE SUMMARY

Introduction

The stereotype regarding female inferiority in mathematics is widespread in the United States and internationally (Nosek et al., 2009). While some scholars have argued that disparities between boys and girls in average mathematics performance are now negligible (Lindberg, Hyde, Petersen, & Linn, 2010), there continues to be a dramatic underrepresentation of girls and women at the highest levels of mathematics achievement (Ellison & Swanson, 2010, Fryer & Levitt, 2010) and in mathematics related professions (Cheryan et al., 2016). Stereotypes regarding appropriate gender roles emerge very early. Negative gender stereotypes are not limited to mathematics performance, but also include that intellectual "brilliance" is also a male-stereotypes quality. A study completed recently with 96 children by Bian, Leslie and Cimpian (2017) documented that girls as young as six were less likely to say that a member of their own gender was "really really smart" compared to boys of the same age. This dissertation study defines the problem of the "gender gap" as both a disparity in mathematics achievement as well as a parallel disparity between boys' and girls' levels of self-efficacy and sense of belonging in the mathematics domain. It documents this problem at a high-level nationally and internationally and then addresses it through an intervention conducted within in the context of a coeducational, independent middle school in the US.

Purpose of Study

The purpose of this study was to assess the gender gap in mathematics performance and attitudes in a coeducational middle school context and to determine if an all-girls supplementary mathematics class, or "math workshop," could help to close this mathematics gender gap for students in fifth and seventh grades. The study sought to contribute to the literature on single-sex

education by determining if this model could reduce stereotype threat for girls who are high-achieving in mathematics and identify with this domain. Furthermore, this study was intended to address the challenges inherent in translating research into the classroom. Therefore, this narrative includes a detailed description of the challenges encountered during the pilot year implementation and recommendations for practitioners.

Theoretical Alignment

This study is grounded in a social-cognitive framework and uses Bandura's (1986) theory of self-efficacy to understand the gender gap in mathematics achievement. In this theory, there is a triadic reciprocal relationship between a person's behavior, environment, and beliefs such that each bidirectionally affect the others (Bandura, 2011). This is an appropriate framework because it incorporates not only individual factors but also larger social-structural factors such as gender that can operate through psychological mechanisms to influence behaviors. Importantly, this theory also attributes most stereotypic attributes of gender to cultural and not biological differences (Bussey & Bandura, 1999).

Literature Review

There is a great deal of research literature that seeks to understand the gender gap in mathematics performance at the high-end of achievement. A large portion of it attempts to determine if there are any biological differences between boys and girls that could mediate mathematics achievement (Geary, Saults, Liu, & Hoard, 2000; Halpern et al., 2007; Jansen, Zayed, & Osmann, 2016; Voyer, Voyer, & Bryden, 1995). At this point, biological and neuroscientific research have found little evidence between the brains of girls and boys that would reliably explain differences in learning (Eliot, 2013). More promising is work documenting the powerful influence that sociocultural influences can have on individual's

performance. In particular, research on psychological mechanisms of stereotype threat, the fear of confirming a negative stereotype about one's group, is helpful in making sense of the gender gap in mathematics. A body of research literature indicates that girls who are high achieving in mathematics may underperform in coeducational settings because gender is more salient when negative stereotypes are primed (Neuville & Croizet, 2007; Steele & Ambady, 2006).

Although research on the potential benefits of single-sex classrooms in general has been equivocal (Pahlke, Hyde, & Allison, 2014), more targeted research on the advantages of all-girls classrooms for male-stereotyped subjects indicate that this approach could be helpful in promoting girls confidence and performance in mathematics by reducing the salience of gender as an identifying variable and lessening the negative effects of stereotype threat (Eisenkopf, Hessami, Fischbacherand, & Ursprung, 2014; Kessels & Hannover, 2008; Picho & Stephens, 2012).

Methodology

This study used a mixed-method approach. The study had a sequential design with a primary emphasis on quantitative data. Standardized test data and survey responses were gathered during the needs assessment to document the size and scope of the problem. During the intervention assessment, a pretest-posttest design was used with multiple comparison groups. Random assignment to the intervention was not possible because of the ethical constraints of educational fieldwork conducted in a school (Rossi, Lipsey, & Freeman, 2004).

The growth in CTP-4 Mathematics achievement scores, the annual standardized mathematics test, and survey responses regarding mathematics self-efficacy and sense of belonging for student who participated in the intervention were compared with students who did not receive the intervention. In addition, the change in scores for girls who participated in the

intervention were also compared with boys who participated in the intervention. In a second phase, qualitative data in the form of interviews conducted with middle-school girls who participated in the intervention was used to help understand the quantitative results and to strengthen reliability and validity by providing multiple sources of data that could be used for triangulation (Creswell & Clark, 2011).

Results

Analysis of CTP-4 mathematics tests results suggested that the math workshop program was effective in reducing the gender gap in mathematics test scores for some students. A 3-way between groups ANOVA was conducted to compare the main effect of gender (boy, girl), track (on-level, advanced) and intervention (treatment, control) on the change in CTP-4 Mathematics achievement scores. There was a significant 3-way interaction, F(1)=7.428, p=.007, $\eta 2$=.051. Specifically, the results indicate that the all-girls "math workshop" intervention helped to close the gender gap between high-ability boys and girls who received the intervention, with high-ability girls gaining more scaled score point on the annual standardized mathematics test than high-ability boys, *t*(57)=-2.36, p=,022, d=.60. The benefits for students of average ability were more equivocal, and did not reach statistical significance.

Analysis of survey responses indicated that students who participated in the math workshop intervention saw a greater increase in sense of belonging scores compared to student who did not receive the intervention, *t*(200)=2.299, p=.023. However, this result was not significant by gender. A 3-way ANOVA for the main effects of gender (boy, girl), track (on-level, advanced) and intervention (intervention, control) on students' self-efficacy found no statistically significant effects. Thus, the intervention appears to have been mediated more by a change in students' sense of belonging in mathematics than their self-efficacy.

Qualitative data in the form of interview transcripts were used to help make sense of the quantitative findings. Of the 12 girls interviewed, nine of them distinguished math workshop from their mathematics class as being more "fun," which they attributed to a variety of reasons including the all-girls environment, the opportunity to work with friends, the reduced focus on tests, and more open-ended material. Of these, the most dominant theme was the importance of friends. Eight of the 12 students interviewed mentioned working with peers as something that helped them feel more confident and more comfortable in mathematics class.

Findings and Discussion

The results of this mixed method study at a coeducational middle school make a contribution to the literature on the gender gap between high-ability boys and girls in mathematics self-efficacy, sense of belonging and achievement. It provides insights into the potential challenges and benefits of implementing a supplementary all-girls mathematics class in a coeducational school. There are several main findings from this study. 1) Supplementary all-girls mathematics classes may be beneficial for certain populations. In particular, this study provides support for the hypothesis that all-girls classes may improve achievement for girls of high-mathematics ability and identity. 2) Mathematics classes or supplemental mathematics activities that are single-sex may increase middle-school students' sense of belonging regardless of gender 3) Sense of belonging in girls may have been influenced by the "friend effect," or girls' reported increase in confidence and enjoyment when solving mathematics problems with their friends.

CHAPTER 1: THE GENDER GAP IN MATHEMATICS ACHIEVEMENT: A REVIEW OF CURRENT LITERATURE

Introduction

High-ability girls (*girl* and *boy* will be used to refer to a student's gender identity as separate from their assigned sex at birth) are underperforming in mathematics achievement compared to their male peers (Fryer & Levitt, 2010; Penner & Paret, 2008; Robinson & Lubienski, 2011). This gender gap in mathematics achievement has been well documented both in the United States and internationally (Organisation for Economic and Co-operative Development [OECD], 2015; Penner & Paret, 2008; Robinson & Lubienski, 2011). The gender gap has been measured as early as first grade and persists as women continue to be underrepresented in mathematics in higher education and mathematics-related professions (Ellison & Swanson, 2010; OECD, 2015; Penner & Paret, 2008; Stoet, Bailey, Moore, & Geary, 2016; Wang & Degol, 2016).

These gender differences in mathematics achievement are echoed in a stark underrepresentation of women in higher-level mathematics. At the graduate level, women currently receive only 29% of doctorates in mathematics and statistics (National Center for Education Statistics, 2014). In 2015, after 64 years, Maryam Mirzakhani became the first woman to receive the Fields Medal, widely considered the highest honor bestowed in the field of mathematics (International Mathematics Union, 2014). Both biological and sociocultural explanations have been explored to explain the gender gap (Ceci, Williams, & Barnett, 2009; Kane & Mertz, 2012; Wai, Cacchio, Putallaz, & Makel, 2010; Wang & Degol, 2016). At this point, research on biological differences affecting cognition between boys and girls is inclusive (Eliot, 2011). More compelling is the large body of research that suggests girls and women may

underperform in part because of differences in societal beliefs and expectations regarding boys' and girls' abilities in mathematics (Kane & Mertz, 2012; Wang & Degol, 2016).

Surveying peer reviewed research published between 2000 and 2017; this literature review will explore current explanations for the gender mathematics achievement. It will begin with an overview of the evidence that there is a substantial gender gap in mathematics performance and that this gender gap is more extreme between high-ability boys and girls. Next, it will present research on biological explanations for this difference in mathematics achievement, and discuss why this argument regarding innate gender differences is not currently a persuasive explanatory model. Finally, it will outline the theoretical framework of social cognitivism (Bandura, 1986) that is used for this research. It particular, it will examine current research pertaining to two constructs—self-efficacy and sense of belonging—and the means by which they may impact girls' mathematics achievement.

Limitations of Literature

Despite large data sets of hundreds and even thousands of individuals, the majority of studies that establish the scope of the gender gap in mathematics do not use a randomized experimental design and therefore cannot make any causal inferences (Fryer & Levitt, 2010; OECD, 2015; Penner & Paret, 2008; Robinson & Lubienski, 2011). Due to constraints of working with children, many experimental studies that have the potential for stronger causational inference are largely conducted with older students in high school and college (Good, Rattan, & Dweck, 2012; Kiefer & Sekaquaptewa, 2007; Nosek et al., 2009). Another limitation of many of the studies on the mathematics gender gap is the use of data from tests such as the Early Childhood Longitudinal Study Kindergarten (ECLS-K), which cannot adequately distinguish among students at the highest percentiles of achievement. Thus, attempts to assess the disparity

in mathematics achievement among top students are often limited to mathematics competitions and other settings of self-selecting students (Ellison & Swanson, 2010).

The Mathematics Gender Gap in International Contexts

In most developed countries, girls score lower than boys on mathematics achievement tests and the difference is larger among the highest achieving students (Stoet et al., 2016). On the 2012 Programme for International Student Assessment (PISA), an international test that explicitly aims to assess mathematics problem-solving abilities, 15-year-old boys scored an average of 11 points higher than girls on problem-solving across all 65 countries and economies that participated; among the top 10% of students, the gender gap averaged 20 points (OECD, 2015). Conversely, in the United States, girls tend to earn higher grades than male students and *average* differences on standardized assessments tend to be small (Ceci, Ginther, Kahn & Williams, 2014; Lindberg et al., 2010; Wang & Degol, 2016). However, the data tell a different story at the right tail of the achievement distribution among the highest performing students in mathematics. The research literature does not have an agreed upon definition for "right tail," and it has been described in different research studies as including a cohort from the top 10 percent of students to the top 0.1 percent of students (OECD, 2015; Penner & Paret, 2008). Whatever cut-off is used to demarcate "high-ability" students, the pattern remains the same: the ratio of boys to girls increases dramatically at the higher percentiles of achievement (Ellison & Swanson, 2010; OECD, 2015; Wang & Degol, 2016).

In order to understand what cultural factors could mediate the gender gap in mathematics achievement, Fryer and Levitt (2010) compared the ECLS-K data with findings from the 2003 PISA and 2003 Trends in International Mathematics and Science Study (TIMSS). The TIMSS is an international assessment of mathematics and science knowledge given to fourth and eighth

grade students around the world. In 2003, there were 47 participating countries. Their analysis focused on kindergarten through fifth grade. They documented a positive relationship between gender equality as measured by the World Economic Forum's Gender Gap Index (GGI) and a smaller gender gap in mathematics. The GGI takes into account several factors related to women's wellbeing including economic participation and opportunity, educational attainment, health and political empowerment (Hausmann & Tyson, 2015). In the most gender-equal countries, the gender gap disappeared altogether. However, when countries that were included only in the TIMSS study and do not participate in the PISA were added to the data set, this positive relationship between GGI and the gender gap disappeared. The change resulted from the fact that these additional countries—largely Middle Eastern countries with low scores on the GGI like Bahrain, Jordan, and Iran—had no gender gap in test scores.

The authors hypothesize that this equality in test scores may be due in part to the exclusive use of single-sex schooling during secondary school in these countries. A regression analysis of these countries suggested that in countries with a high-level of sex-segregated school female students are doing better and male students are doing worse. Fryer and Levitt (2010) propose that coeducational schooling may be a prerequisite of gender inequality in mathematics achievement, although they acknowledge that the tendency for these countries to be Islamic is a cofounding variable. Regardless of the causes, the existence of any countries in which the gender gap is nonexistent lends powerful evidence to the theory that the gender gap is sociocultural in nature and not biologically based.

The Gender Gap in the United States

The gender gap in mathematics achievement has been documented in the United States as early as first grade (Penner & Paret, 2008). Previously, literature recorded the emergence of a

gender gap in middle school or as late as high school (Hyde et al., 1990). More recently, the gender gap in mathematics has been shown to exist much earlier in students' lives, and potentially even before students enter formal schooling. Penner and Paret (2008) used data from the Early Childhood Longitudinal Study kindergarten (ECLS-K) Class of 1998-99, which provides educational information on a nationally representative cohort of students from Kindergarten through fifth grade. The sample size was large, with more than 11,000 students included. The authors used quantile regression models to look at gender differences across the distribution. They found that gender differences were present in kindergarten, when girls did better at the bottom of the distribution and boys did better at the top of the distribution. However, by the end of the third grade, the girls were no longer out-performing males in the bottom quartile and the boys' advantage extended to most of the distribution. Although these early differences were small, about 0.15 standard deviation units, the authors noted that they are significant because small advantages may compound over time into much larger differences in achievement. This is particularly true in mathematics because of its cumulative nature. The authors also documented differences in the gender gap across racial groups and also by parents' educational background. For example, for the Latino population the sample female students actually entered kindergarten with an advantage at the top of the achievement distribution, and the male advantage was worse at the top of the distribution in families with parents who held a college or advanced degree. Based on these findings, Penner and Paret (2008) proposed that cultural factors are more likely to explain the gender gap than biological ones.

A second analysis of the ECLS-K data set collected by the U.S. Department of Education revealed a similar pattern. Robinson and Lubienski (2011) used a subset of ECLS-K data from 7,075 students, all of whom ranked high in English proficiency and who had complete

assessment information, to track changes in mathematics achievement from Kindergarten through eighth grade. Consistent with the work of Penner and Paret (2008), the authors found that boys showed an advantage as early as first grade at the top end of the distribution (90^{th} percentile), and that over time this advantage expanded such that boys were also performing better at the 50^{th} and 10^{th} percentiles by third grade and fifth grade, respectively. By the spring of fifth grade, the gap between boys' and girls' achievement was between 0.22 and 0.30 standard deviation units at each of the percentiles examined. There appeared to be some narrowing of the gap at the 90^{th} percentile between fifth and eighth grade, such that only 37% of the top 1% of eighth grade students were girls—a notable increase from 15% in Kindergarten, though still leaving girls a significant minority of the top mathematics performers. Robinson and Lubienski (2011) suggested that this reduction in the gender gap might be due to gender-focused intervention programs targeted at middle-school girls or perhaps to the increasing importance of homework, which girls tend to report spending more time completing. The authors concluded that while these differences in mathematics achievement are concerning, they should not be considered inevitable. Given that there was some closing of the gap in their particular data set, the authors urged further research into interventions that might further close the disparity. They suggested that mathematics-focused interventions to help female students may be more effective if they are targeted at elementary school children.

Fryer and Levitt (2010) also used data from ECLS-K, a sample of approximately 20,000 children from 1,000 schools, to examine the emergence of the gender gap at the upper tail of mathematics achievement. On entry into kindergarten, females made up 45% of the top fifth percentile in mathematics performance, but by the end of the fifth grade females constituted only 28% of the top fifth percentile. These results provide persuasive evidence of large gap in

mathematics performance between high-ability males and females that develops early in elementary school. This gender gap was found in every demographic of society, but was larger for students who attended private school, had highly educated parents, or a mother who worked in a mathematics-related occupation. These findings were counterintuitive, as these are all factors that might be considered helpful in promoting girls' success. The authors further discarded previous explanations for the gender gap, including parental expectations, biased tests, or investment in mathematics activities.

In a study of the gender gap that examined geographic differences, Pope and Sydnor (2010) used data from the National Assessment of Educational Progress (NAEP), a set of standardized tests that are given to students in grades 4, 8, and 12 across the United States. The NAEP is given to a large proportion of students nationally, and does not aim specifically to distinguish among students at the highest end of the distribution. In an analysis of eighth grade mathematics scores there were only slight differences in the mean score for girls and boys. However, a gender gap emerged among students scoring in the top 5%.

Pope and Sydnor (2010) analyzed the NAEP test scores by geographic region. In every state other than Hawaii, boys outperformed girls at the highest levels of mathematics and girls outperformed boys at the highest levels of reading. At the same time, there was geographic variation in the degree of disparity, and those states that had a large gender gap on one test tended to have a large gender gap on the others. The authors suggested that there may be geographic areas that adhere more to gender stereotypes than others. New England had the lowest average disparity between gender performance on both mathematics and reading, while the east south central region had the highest differences by gender. Children in regions with larger gender gaps were also more likely, in an unrelated survey from 1992, to have agreed with

the statement “math is for boys” (Pope & Sydnor, 2010, p. 107). This result provides support for the hypothesis that the gender gap is correlated with indicators of gender equality in a society.

The Gender Gap at the “Right Tail”

Gender differences in mathematics achievement are larger at the highest levels of achievement, and are thus more visible on highly competitive tests that have a higher “ceiling” for scoring. Much of the data that are used to examine the gender gap is national achievement data from elementary school such as Early Childhood Longitudinal Study Kindergarten (ECLS-K) (Penner & Paret, 2008), and therefore does not distinguish well among students who score in the top percentiles. Instead, researchers interested in this “right tail” have used data from college-entrance exams and mathematics competitions (Ellison & Swanson, 2010). Between 2006 and 2010, male students outnumbered female students at the top .01% of the distribution at a ratio of 4:1 and 3:1 on the SAT and the ACT respectively (Wang & Degol, 2016). The male to female ratio of top-performing students is even more extreme, often as high as 10:1, on high-level mathematics assessments such as the American Mathematics Competition (AMC), which is designed to be a challenging mathematics competition with a “high ceiling” that distinguishes among already high-achieving mathematics students (Ellison & Swanson, 2010). On the AMC twelfth competition given November 2016, all of the top 35 scoring students were boys (i.e., among students who identified gender; Mathematical Association of America, 2016).

Multiple research studies confirm the finding of the 2015 PISA report that the gender gap is more extreme at the highest percentiles of performance. In one influential study, Ellison and Swanson (2010) document a significant gender gap on mathematics achievement at the highest percentiles of achievement using statistical analyses of the American Mathematics Competition (AMC) since 1950. Given to about 225,000 high-school students nationwide, this elite contest

provides more effective data than the SAT for distinguishing among the very top level of mathematics performers. The difficulty-level of the SAT results in ceiling-effects that do not adequately represent the achievement differences among top performing students. The authors found that on the AMC the gender gap widens at higher percentiles. Above the 99th percentile, the gender gap reached a male-female ratio of 10:1 in the 2007 AMC competition. This was a pervasive pattern across the country, although there was some variation. One of the key findings of the study was that the top-performing males are drawn from a wider pool of schools than the top-performing females. The authors suggested that there may be only a small number of schools that are effectively supporting females in reaching high-levels of problem-solving achievement, leaving a large population of talented girls who are not fully developing their potential. Ellison and Swanson (2010) suggest that girls may be more compliant with authority figures and less likely to ask for and attain the special accommodations that may be required to prepare them for an elite competition such as the AMC.

Biological vs. Sociocultural Explanations

Both biological and sociocultural explanations have been suggested for the gender gap in mathematics achievement. Thus far, evidence for an innate, biological difference in mathematics ability has been inconclusive (Ceci et al., 2009; Eliot, 2011; Halpern et al., 2007; Kane & Mertz, 2012; Wang & Degol, 2016). The biologically-based explanations are founded primarily on the *greater male variability hypothesis* that boys and men are overrepresented at both the high end and low ends of achievement on mathematics tasks (Geary et al., 2000; Halpern et al., 2007; Jansen et al., 2016; Voyer et al., 1995).

Specifically, a great deal of research has focused on gender differences in mental transformation, the ability to imagine and rotate two and three-dimensional objects (Jansen et al.,

2016). Although males do tend to outperform females on tests of spatial rotation, the gender differences are moderated by the conditions of the task, with the gap in scores diminishing when time constraints are removed (Voyer, 2011). The results of mental rotation tasks also differ by country. A recent study by Jansen et al. (2016) comparing the mental rotation abilities of college students in Oman and Germany (n=239) found that boys (M=10.15) outperformed girls (M=8.10), but there was an even greater difference between students from Oman (M=7.32) and German (M=10.92). The authors propose that the differences could be due to educational factors or participation in spatial activities as a child. The large moderating effect of country of origin suggests that mental rotation ability is not completely innate or biologically determined.

Other research into gender differences in mathematics performance has investigated the possible role of testosterone and other sex hormones (Ceci et al., 2014; Quaiser-Pohl, Jansen, Lehmann, & Kudielka, 2016). A recent study of mental rotation performance in 109 children ages 9 to 14 found no significant gender differences in hormone levels for either testosterone or estradiol, and hormone levels had no significant interaction with reaction time or accuracy (Quaiser-Pohl et al., 2016). The authors also did not find that boys had better accuracy than girls in the study, only that they tended to be slightly faster in responding. In a recent review of research into the effects of sex hormones on cognition, Eliot (2011) writes “all of this evidence tell us that circulating hormones have little, if any, effect on human cognition” (Eliot, 2011, p. 374). This author further points out that this explanation does not hold for elementary-school children who have no little differences in sex hormones because they are prepubescent.

Biological explanations for differences in mathematics achievement are also complicated by research that demonstrates girls tend to do better in mathematics achievement when they are evaluated based on school grades instead of standardized tests (Egorova, 2016; Voyer & Voyer,

2014). In a meta-analysis of 392 samples, Voyer and Voyer (2014) estimated that girls outperform boys based on grades in middle and high school in every subject area, including mathematics. The effect size (d=.069) for mathematics was the smallest of the courses analyzed, but still favored girls. Although the effect sizes are small, the accumulated documentation of this female-advantage is quite striking, particularly when it contradicts the trend that boys are outperforming girls on standardized tests of mathematical achievement. The study, however, focused on mean differences and did not specifically look at gender differences at the right tail. Egorova (2016) conducted an original study comparing mathematics performance by gender of Russian teenagers. The authors found that girls did better based on school grades for both algebra (d=.33) and geometry (d=.41). A different result was found from analyzing standardized test scores. Just as in the United States, boys had a very small advantage on average (d=.05), but Russian boys tended to outperform Russian girls at the very high-end of mathematics achievement, which the authors defined as more than two standard deviations above the mean. The authors conclude that there are no easy explanations for these contradictory results and propose that they are most likely due to complex mechanisms involving self-concept and motivation.

Another reason to question the biological explanations for the gender gap is the fact that measured gaps of mathematics achievement have changed over time. In the United States, for example, the gap between the average male and female students on the mathematics section of the SAT has shrunk from 40 points to 33 (Ceci et al., 2009). At the high end of the distribution, there has also been a narrowing of the gap between boys and girls. The male to female ratio at the top 0.01% of mathematics achievement on the SAT was 13:1 in the early 1980s and then dropped to approximately 4:1 in the 1990's where it has remained (Wai et al., 2010).

Although biological factors cannot be dismissed in understanding the gender differences favoring boys among top mathematics achievers, the fact that the gap in mathematics performance is not present in every country, and that it has narrowed over time in the United States, strongly suggests sociocultural causes play a significant role (Ceci et al., 2014; Wai et al., 2010). If differences in innate ability between the sexes were a primary causal factor, one would expect there to be relative consistency across countries and time (Ceci et al., 2009). Instead, although the gender gap at the right tail is well documented, it does not appear immutable and varies by culture, time, and socioeconomic status (Ceci et al., 2014; Kane & Mertz, 2012, Voyer, 2011).

Theoretical Framework

While biological explanations have proved inadequate, researchers seeking to understand the gender gap have increasingly looked to sociocultural factors (Kane & Mertz, 2012; Wang & Degol, 2016). In a recent report on a comprehensive research project on gender equality, the OECD concluded: "gender disparities in performance do not stem from innate differences in aptitude, but rather from students' attitudes towards learning and their behavior in school…" (2015, pg. 5). However, there is still much to be learned about the relevant constructs that contribute to boys' and girls' different attitudes towards mathematics or how these attitudes might mediate achievement, particularly for boys and girls at high levels of mathematics achievement.

This literature review applies a social cognitivist lens to the problem of the gender gap in mathematics. Social cognitive theory is grounded on the idea that human agency, the ability to exert control over one's environment, is central to the human experience (Bandura, 2001). Further, people construct their expectations of future outcomes based on observations of how

previous actions have resulted in past outcomes (Bandura, 1986). Individuals are then capable of reflecting on the accuracy of their own predictions as well as the effects of their own actions and other peoples' actions (Bandura, 2001). From these reflections, people develop beliefs about their ability to exert control over their environment and future events, which Bandura (1986) defines as self-efficacy. Without self-efficacy, people have little incentive to persevere when confronted with difficulties (Bandura, 2001). Efficacy beliefs can modulate how much effort people expend on a given task and shape whether they interpret failures as motivating or discouraging (Bandura, 2001). Social cognitive theory represents the relationship between personal beliefs, behavior, and environment as triadic reciprocal causation, (Figure 1) in which each factor affects the others bidirectionally (Bandura, 2001).

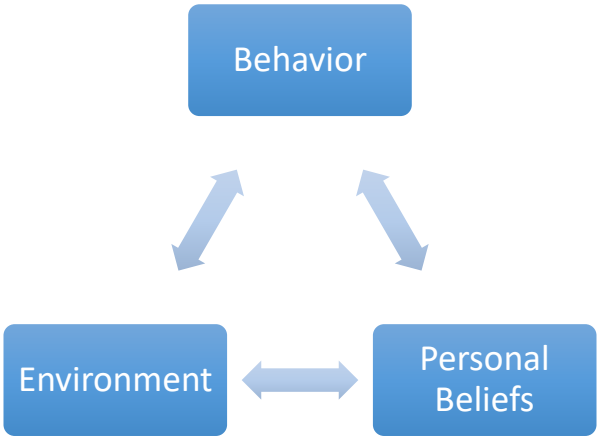


Figure 1. Bandura's theory of triadic reciprocal causation.

Social cognitive theory can also be used to understand efficacy at a group level. Groups of people can have a shared sense of agency or lack of agency that is determined by complex interactions among the group members, and is not equal to the sum of the efficacy beliefs of individual members (Bandura, 2001). In addition, sociostructural factors such as socioeconomic status, education, and family structure operate through psychological mechanisms to produce

behaviors (Bandura, 2001). Social cognitive theory proposes that most of the stereotypic attributes of genders are due to cultural influences instead of biological differences (Bussey & Bandura, 1999).

This perspective argues that gender roles and behaviors result from societal influences both within the family and in larger social systems of everyday life. More specifically, this theory suggests that gender development is affected by three major methods of influence: modeling, enactive experience, and direct teaching (Bussey & Bandura, 1999). With modeling, gender-linked information is gathered from the immediate environment by observing parents, peers, and other role models. Enactive experience refers to gathering information from others' reactions to one's behavior. Generally, gender-linked behavior is highly socially promoted. The third method of gender influence is through direct teaching about what is appropriate for each gender (Bussey & Bandura, 1999). To fully understand a female students' sense of efficacy in mathematics, it is necessary to examine agency beliefs at both an individual and group level.

Causal Model

A proposed model of the etiology of the gender gap between boys and girls is presented below, showing how the various relevant constructs relate to one another. This model suggests that the coeducational gender composition of a mathematics classroom can increase a girls' awareness of her gender, or gender salience, and activate negative stereotypes about girls' ability in mathematics (Neuville & Croizet, 2007; Steele & Ambady, 2006). This experience of stereotype threat, the fear of confirming a negative stereotype about one's group, can reduce a girls' self-efficacy and sense of belonging in the classroom and can lead to reduced performance on achievement on standardized mathematics tests (Else-Quest, Hyde, & Linn, 2010; Good et al., 2012; Steele, 1997).

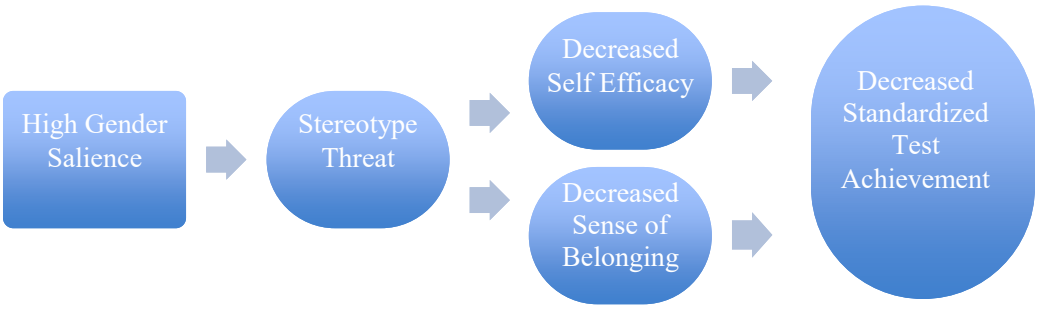


Figure 2. Diagram of causal model.

Gender Salience and Stereotype Threat

There is a growing body of literature on the role that gender salience plays in priming negative gender stereotypes about women in mathematics (Neuville & Croizet, 2007; Steele & Ambady, 2006). Gender salience is defined as the degree to which gender identity is activated and emphasized (Neuville & Croizet, 2007). Increasing gender salience can prime gender-stereotyped views about mathematics (Steele & Ambady, 2006) and increase the detrimental effects of stereotype threat on mathematics performance (Neuville & Croizet, 2007). The research on gender saliency and stereotypes reveals that even individuals who disavow stereotypical views about men and women will often show implicit stereotyping (Kiefer & Sekaquaptewa, 2007; Steele and Ambady, 2006). Implicit stereotypes refer to a cognitive association between a particular social group and stereotypical attributes (Kiefer & Sekaquaptewa, 2007). Implicit and explicit stereotypes are not strongly correlated and may need to be considered as separate constructs (Kiefer & Sekaquaptewa, 2007). One limit to the existing research on gender salience and stereotypes is that much of it has been conducted in research labs and not classroom environments.

In one experiment conducted by Hilliard and Liben (2010), preschool children ages 3-5 (n=57) completed measures using the Preschool Occupation, Activity, and Trait-Attitude

Measure (POAT-AM) of gender attitude, intergroup bias, and personal preference. The children were then studied as they took part in one of two conditions. In the gender low-salience condition, the preschool teacher avoided making gender explicit. In the high-salience condition, the teacher highlighted gender by using gender-specific language, labeling groups with gender terms, and organizing some activities by gender. The researchers predicted that children in the high-salience condition would demonstrate an increase in gender stereotypes, an increase in in-group bias (preference for their own gender), and an increased interest in activities stereotyped for their own gender. Gender-linked interests were measured using the Preschool Occupation, Activity, and Trait-Attitude Measure (POAT-AM). Results indicated that children in the high salience group did show increased stereotyped attitudes and increased avoidance of the opposite sex. Although the study has only a moderate sample size, it conveys the power that subtle cues about gender salience can have on students. In this case, the students were very young, and using inclusive language and activities could mitigate gender salience. It is not yet known how this approach would work for adolescent students in a coeducational setting. In addition, this is one of the first studies of gender salience research in a classroom setting.

In a study of slightly older children, Neuville and Croizet (2007) investigated whether raising gender salience would affect the performance of 7- and 8-year-old girls solving a mathematics problem. They also examined the role that task-difficulty played in performance. Using the theoretical model of stereotype threat (Steele, 1997), they predicted that activating gender stereotypes about females would inhibit performance on difficult mathematics tasks but not easier ones. In the experiment, boys and girls (n=79), were assigned to either a gender-activation group or gender-non-activation group. In the activation group, girls colored a picture of a girl holding a doll and males colored a picture of a boy holding a ball. In the non-activation

group, both boys and girls colored a landscape. All students were then asked to solve 7 problems, five of which were considered easy and 2, which were more difficult.

The results indicated that girls in the gender-activation group underperformed on the difficult questions compared to girls in the non-activation group. Interestingly, girls in the gender-activation group also did marginally better on the easy questions. There was no significant effect for males. Neuville and Croizet (2007) concluded that the study confirms that gender salience can have a detrimental effect on girls' classroom performance in mathematics when they are faced with challenging problems. This is one of the first studies to explore how the impact of gender salience on mathematics performance may differ depending on the difficulty of the problems involved. This study may also shed light on why high-ability girls do not perform as well on very challenging mathematics competitions such as the AMC. In these cases, stereotype threat may be even more detrimental and result in high-ability girls' underperformance. The authors hypothesize that a threatening environment can counter-intuitively help girls on easier problems because of heightened arousal. However, the authors also acknowledge that easy problems may simply be too easy to be sensitive to any impairment because they require so few cognitive resources.

The effects of stereotype threat may function subconsciously. In a study with college-age women (n=46), Steele and Ambady (2006) found that women who were subconsciously reminded of the category "female" later expressed attitudes towards mathematics that were more consistent with female gender stereotypes. The participants were first shown a string of words for brief "flashes" that subliminally primed the category of "female" or "male." Next, the participants were asked to rate a variety of activities related to arts and mathematics as more or less pleasant. The results showed that females who had been primed with "female" indicated a

greater preference for arts activities over mathematics. This pattern was not seen for women who had been primed with "male." This study expands previous research by indicating that even when women are primed subconsciously about their gender identity it can have a detrimental effect on their attitude towards mathematics. In a second study, Steele and Ambady (2006), asked female college students (n=35) to fill out a form that either primed their gender by asking about their sex, their living condition (coed or single-sex), and the advantages and disadvantages of each living arrangement, or did not prime their gender identity by instead asking questions about telephone service. Using the same questionnaire from study 1, paired t-tests revealed that female students in the primed condition expressed a personal preference for arts activities over mathematics while those in the neutral condition did not.

Research into stereotype threat has also attempted to measure the effects of negative stereotypes that operate below conscious awareness. Kiefer and Sekaquaptewa (2007) specifically examined how implicit gender stereotypes interact with mathematics performance on a preconscious level. Sixty-three female undergraduate students enrolled in a first-year calculus class participated. Participants completed an Implicit Association Test (IAT) to measure gender stereotyping and mathematics. The test used reaction time to measure association between categories. The participants were also asked to answer survey questions with differing levels of agreement to assess their level of gender identification and explicit stereotypes. The researchers found that students who rated themselves lowest for gender identification and had lower implicit gender stereotypes did better on the final calculus exam and expressed more interest in pursuing a mathematics-related career. These results remained true after controlling for SAT scores and previous performance. The authors propose that both conditions (low gender identification and low implicit stereotypes) are necessary to protect against stereotype threat. This study builds

upon previous research in implicit gender stereotypes by combining it with the concept of gender self-identification. Those women who most identify as female may be at a greatest risk of stereotype threat because they perceive that the stereotype is self-relevant (Kiefer & Sekaquaptewa, 2007). Interestingly, the authors found no correlation between explicit stereotyping and performance. It may be that implicit stereotypes are more powerful in understanding women's underperformance in mathematics.

In one of the largest studies of implicit gender stereotypes, Nosek et al. (2009) analyzed results from the Implicit Association Test (IAT) for adults (median age= 27) from 34 different countries. The IAT measured how quickly participants were able to categorize items. It asked participants to pair items that represent *female* (e.g., she, girl) together with items representing *liberal arts* (e.g., arts, history) and to pair *male* items (e.g., he, boy) with *science* items (e.g., physics, chemistry)—or to do the reverse (*male*/*liberal arts* and *female*/*science*). The majority of people were able to categorize words faster in the first condition (*female/liberal arts* and *male*/*science*). The authors interpreted this result as an implicit gender-science stereotype (Nosek et al., 2009).

In this study, the authors used regression analysis to compare the results of the IAT with results from the 2003 TIMSS (Trends in International Mathematics and Science Study) for eighth grade students in 34 countries (n=298,846), with the largest portion of the sample coming from the United States (248,306). They found a strong positive relationship (r=.060) between national results on the IAT and the gender gap in eighth grade science performance. The authors found a similar relationship between mathematics scores and gender-science stereotypes. This relationship was still present after controlling for societal gender inequality. The authors propose that both eighth grade test-takers and participants in the IAT are influenced by social-cultural

norms regarding gender in which science and mathematics are associated with being male. They further suggested that interventions to address the gender gap in mathematics and science achievement must address embedded implicit stereotypes. The large sample size of this study makes it a powerful piece of research in establishing the important relationship between implicit stereotypes about gender and mathematics achievement outcomes.

Recent research suggests that implicit gender stereotypes and children's identification with these stereotypes begin at a young age (Cvencek, Meltzoff, & Greenwald, 2011). In an article by Cvencek et al. (2011), the authors investigate the gender-stereotypes and mathematics self-concepts of elementary-age students using both implicit and explicit measures. The study sample included 126 females and 121 males in grades 1-5. The students completed both an Implicit Association Test (IAT) and self-report measures. As expected, boys tended to associate *me* and *math* more than girls did on both measures. Boys also associated *math* with *own gender* more than girls did, indicating that gender stereotypes about mathematics may already be present as early as first and second grade, around the same age when significant differences in mathematics achievement are becoming apparent (Penner & Paret, 2007). The authors, both cognitive psychologists, hypothesize that the differences in gender identification are due to a combination of societal influences. This study makes an important contribution to the literature on the emergence of early-childhood gender stereotypes about mathematics. It helps to establish that gender stereotypes related to mathematics learning may be present almost as soon as formal schooling begins.

Mathematics Self-Efficacy

Mathematics self-efficacy is an individual students' perception that he or she can successfully complete a mathematics problem (Bandura, 1986). Self-efficacy is a key mediating

factor for achievement, even after controlling for previous performance (Hall & Ponton, 2005; Pajares & Miller, 1994). Bandura (1997) proposed that self-efficacy is determined primarily from four sources: one's own mastery experiences; vicarious observation and comparison with others, social persuasion of parents, teachers and peers; and emotional and physiological states. A body of research literature suggests that girls' tend to have lower self-efficacy in mathematics than boys (Herbert & Stipek, 2005; OECD, 2015; Simpkins, Davis-Kean, & Eccles, 2006).

In one study aimed at developing a self-efficacy in mathematics scale for middle school students, Usher and Pajares (2009) used multiple regression analyses to test Bandura's (1986) theory that students develop their self-efficacy beliefs by interpreting information about their own abilities and performance. The authors found that middle school students' (n=824) own academic experience was the most consistent predictor of students' beliefs about self-efficacy, explaining more than 20% of variance, but that vicarious experience explained 16% of the variance, and both social persuasion and emotional states were also contributing factors. The results were invariant across gender. This study substantiates the claim that self-efficacy beliefs are formed from a variety of interacting sources including self, peers, and parents.

The 2015 OECD report on gender equity found that on average, girls who participated in the 2012 PISA expressed lower self-efficacy about mathematics. Furthermore, these negative attitudes were correlated with a lower performance on the test, equivalent to nearly a whole year of school. Furthermore, self-efficacy scores were more predictive of performance for high-performing students than low-performing students. A difference of one unit on the *index of mathematics self-efficacy* was associated with a 43 score-point difference for students among the lowest 10% of participants but with a 53-point difference in performance for students in the highest 10% of participants (OECD, 2015). Therefore, self-efficacy appears potentially even

more significant in predicting outcomes for high-ability female students than students of average mathematics achievement, offering one potential explanation for the widening gap between boys and girls at the right tale.

Several studies have compared boys' and girls' self-efficacy in mathematics in elementary school (Herbert & Stipek, 2005; Simpkins & Eccles, 2006). In a longitudinal study of 300 children's beliefs from kindergarten or 1st grade through fifth grade, Herbert and Stipek (2005) asked children to report their level of competency in literacy and mathematics on a five-point Likert-type response scale. They found that girls (M=3.91) rated their abilities in mathematics lower than boys (M=4.11) beginning in third grade, despite the fact there was no difference in the teacher's assessment of boys' and girls' mathematics ability or corresponding difference in mathematics achievement. At the same time, girls tended to outperform boys in literacy tests but did not rate their own abilities in literacy higher compared to boys. The researchers conclude that there may be cultural forces that propel girls to be more modest and less self-confident than boys about their academic skills.

Some research supports the hypothesis that not only gender but also level of giftedness may play a role in determining a students' level of self-efficacy in mathematics (Eisenkopf et al., 2015; Hargreaves, Homer, & Swinnerton, 2008; Preckel, Goetz, Pekrun, & Kleine, 2008). German educational psychologists Preckel et al. (2008) investigated the self-concept, motivation, and interest in mathematics of sixth-grade students of gifted and average-ability. The study had a rigorous design due to its random selection of students from a large population. The authors studied 181 gifted and 181 average ability students randomly selected from a larger sample of 2,059 students from across 42 schools in Germany. Giftedness was defined as scoring above the fifth percentile on a nonverbal reasoning subscale of a German Cognitive Abilities test. Each

gifted student was matched with an average ability student of the same gender, from the same school class and with a similar socioeconomic status.

The authors assessed students' beliefs about their competence in mathematics using a questionnaire. Mathematics achievement was measured with a 63-item test based on the concept of mathematical literacy defined by the OECD. Boys scored higher on the mathematics achievements test in both the gifted (mean males score=110.92, mean female score=105.62) and average ability groups (mean males score=103.91, mean females score= 101.39) although there was no difference in the grades the students received from their mathematics teacher. In addition, gifted and average ability girls showed lower self-concept and interest in mathematics. The gender differences in self-concept and interest in mathematics were larger for the gifted students, supporting previous studies that indicate a larger gender gap at the highest percentiles (OECD, 2015; Penner & Paret, 2007). This article is an important contribution to the literature in establishing that lower self-efficacy among gifted female students may be one contributing factor to the larger gender gap in mathematics achievement between high-performing males and females. As the gap increases between the number of boys and girls at the highest level of achievement, the self-efficacy differences parallel this pattern and widens as well.

Sense of Belonging

Another factor associated with higher achievement in mathematics is sense of belonging (Dasgupta, 2011; Good et al., 2012; Smith, Lewis, Hawthorne, & Hodges, 2013). Good et al. (2012) define sense of belonging as a feeling of membership and acceptance. People's behavior and choices are influenced by a need to feel accepted and valued by a community of peers (Dasgupta, 2011). Negative stereotypes imply that certain groups are less welcome or valued (Good et al., 2012; Steele, 2011). Girls and women are susceptible to "stereotype threat," the fear

of confirming the common stereotype that women have poorer mathematics skills than men (Cherney & Campbell, 2011; Else-Quest et al., 2010; Steele, 1997). Girls' exposure to stereotype threat may contribute to lower participation in mathematics activities among female students, even among those who are the highest achievers (Good et al., 2012). Over time, this uncertain level of belonging leads to weaker performance and lower self-confidence in one's abilities in the given domain. This may even be the case when an individual's performance is just as strong as their peers (Dasgupta, 2011). In the United States, girls and women are stereotyped to lack very high levels of mathematics ability, and the majority of girls and women are aware of these stereotypes (Nosek et al., 2009). Previous research indicates that people are more aware of a particular social identity when they feel that social identity is stigmatized (Murphy, Steel, & Gross, 2007). Furthermore, this experience of stereotype threat can decrease an individual's sense of belonging and interest in participation in a given domain (Murphy et al., 2007). Consequently, girls may be even more aware of environmental clues such as the ratio of boys to girls in a room that would indicate that they might not belong there (Smith et al, 2013). Accordingly, when girl are numerically underrepresented in a learning environment they may be more likely to experience stereotype threat, particularly when they identify as high-achieving in the domain (Beilock & Carr, 2005; Dasgupta, 2011).

There has been a recent interest in better understanding the importance of belonging for academic achievement. Good et al. (2012) found that college graduates' sense of belonging in mathematics was a strong predictor of their intention to pursue mathematics classes at the college level. The authors asked 133 participants (56 men and 77 women) who were enrolled in college-level calculus to fill out a 28-item scale measuring their sense of belonging in mathematics. All questions began with "When I am in a math setting," and asked participants to rate their

agreement on a 1-6 Likert scale. A regression analysis indicated that women who reported a higher sense of belonging in mathematics were less likely to report being anxious (regression coefficient of -.70) about mathematics and more likely to report greater confidence in mathematics (regression coefficient .77). This study suggests that sense of belonging is an important construct in understanding female students' attitudes towards mathematics and their level of participation in the mathematics domain. Although the study was conducted with college-aged students, studies with younger children have found that females' report believing that mathematics is a male-stereotyped domain in primary school (Herbert & Stipek, 2005). Further studies are needed on sense of belonging to understand its role for adolescent children.

Mathematics Anxiety

Another key construct that may prove helpful for understanding the links between students' mathematics self-efficacy and their mathematics achievement is mathematics anxiety, which can be broadly defined as a fear and aversion to mathematics (Beilock & Carr, 2005). Mathematics anxiety and self-efficacy are highly correlated, as anxiety is frequently linked to fears of underperformance (Beilock & Carr, 2005). This present research study focuses on self-efficacy and sense of belonging, but key literature on mathematics anxiety is included here as a critical related field. There is important recent work in the field of math anxiety that examines how high-pressure conditions could exacerbate stereotype threat and inhibit optimal performance, particularly among otherwise highly-able individuals (Beilock & Carr, 2005)

Women and appear to be more vulnerable to developing math anxiety because of stereotype threat (Beilock & Carr, 2005; Schmader and Johns, 2003). In mathematics classes, negative stereotypes about women are primed, which can lead to lower performance (Schmader & Johns, 2003). Students with mathematics anxiety tend to enjoy mathematics less, to

underestimate their mathematical abilities and to participate less in mathematics classes (Vukovic, Kieffer, Bailey, & Harari, 2013). Research suggests that mathematics anxiety lowers working memory and achievement on tests (Beilock & Carr, 2005). In this context, working memory is defined as the temporary storage and manipulation of information (Vukovic et al. 2013).

In one of the first studies of mathematics anxiety in young children, Ramirez, Gunderson, Levine, and Beilock (2013) gave 154 students in first and second grades a test of working memory and mathematics achievement. Subsequently, the authors assessed mathematics anxiety using a newly designed scale, with pictures to indicate levels of anxiety. The results of the study suggested that mathematics anxiety disrupted working memory for students with high working memory but not those with lower working memory. The authors argue that mathematics anxiety co-opts working memory, thus forcing students with high working memory to use slower and less-effective mathematical problem solving strategies. Results from a study conducted by Vukovic et al. provide further support for this hypothesis. They found a negative correlation between mathematics anxiety and mathematics performance in 113 second and third graders, with the most pronounced effect seen for children with higher working memory scores. Although counter-intuitive, these results suggest that it is in fact the highest-ability students (i.e., those with high working memory) whose performance is most damaged by math anxiety.

In a related study, Beilock and Carr (2005) established that the same relationship between mathematics anxiety and performance is evident for college-age students. The authors conducted a study of 93 undergraduate students who were divided into two groups, a group with high working memory, and a group with low working memory. The students were asked to solve mathematics problems under high-pressure and low-pressure conditions. The authors found that

students with high working memory performed better compared to the low working memory counterparts on the low-pressure task, but they lost this advantage in the high-pressure conditions. This effect was seen primarily on the tasks that required the working memory. The study used analysis of variance (ANOVA) to establish that individuals with high working memory show statistically significant reduction in problem solving performance under high-pressure conditions. These results support the hypothesis that it is individuals with the highest potential who may be most affected by mathematics anxiety.

Initial research suggests girls may be vulnerable to developing math anxiety if they have a female teacher who is anxious about mathematics. In a compelling study, Beilock, Gunderson, Ramirez, and Levine (2010) argue that female elementary school teachers who are anxious about mathematics may transfer this anxiety to their female students, thereby reducing these students' overall mathematics achievement. The authors surveyed first and second grade teachers and found that at the beginning of the school year there was no relationship between students' mathematics achievement and the teacher's anxiety. However by the end of the school year, the more anxious the female mathematics teacher, the more likely female students were to endorse the stereotypical belief that boys are better at mathematics, and the lower these girls' performances were on mathematics achievement tests. Girls who agreed with the statement "boys are better at math and girls are better at reading" performed worse than girls who disagreed, and worse compared to boys overall. The authors assessed math anxiety using the Math Anxiety Rating Scale (MARS) and student achievement with items from the Woodcock-Johnson III test. The authors used regression analysis to establish a link between teachers' math anxiety and girls' mathematics achievement mediated through gender ability beliefs. Although the sample size was relatively small, 17 teachers and 117 students, this study references a robust

body of literature that supports their findings and found significant effect sizes. Given the fact that greater than 90% of elementary school teachers are female, this early indication that mathematics anxiety may be perpetrated through unconscious transfer of anxiety is extremely powerful. This study also illuminates the cyclic nature of mathematics underperformance among women, whereby low high anxiety impact behavior and performance and in-turn reconfirm negative beliefs about mathematics self-efficacy.

Conclusion

This literature review discussed the scope and size of the mathematics achievement gap. It documented the pervasive finding that males outperform females in mathematics at a small but significant level at the mean level of achievement, and that the size of the achievement gap widens at the highest levels of achievement (Fryer & Levitt, 2010; OECD, 2015; Penner & Paret, 2008). Next, this review proposed a social-cognitive framework for understanding the gender gap in achievement that is built on Bandura's (1986) theory of self-efficacy. It discussed a causal model for this problem, in which females in coeducational classrooms are made aware of their gender and experience the negative effects of stereotype threat (Inzlicht & Ben-Zeev, 2000; Schmader & Johns, 2003). Stereotype threat result in math anxiety and can reduce females' sense of self-efficacy and sense of belonging in the mathematics domain, two constructs that have been linked to mathematics achievement (Else-Quest et al., 2010; Good et al., 2012; OECD, 2015; Steele, 1997). Chapter Three will discuss the scope of the gender gap in a specific school context and the results of an initial needs assessment.

CHAPTER 2: A NEEDS ASSESSMENT AT AN INDEPENDENT SCHOOL

Introduction

As discussed in Chapter 1, there is a significant gender gap in mathematics achievement between high-ability girls and boys (Penner & Paret, 2008; Robinson & Lubienski, 2011). There is evidence from multiple research studies in elementary schools that this gap emerges much earlier than previously thought—perhaps as early as Kindergarten or first grade (Penner & Paret, 2008). The discrepancy in achievement is also extreme at the highest percentiles of achievement (Ellison & Swanson, 2010; OECD, 2015; Penner & Paret, 2008). Scholarly research into the causes of this gap in achievement supports the theory that its primary drivers are sociocultural phenomena and not innate biological differences (Ceci et al., 2009; Kane & Mertz, 2012; Wang & Degol, 2016).

An initial review of the academic literature on attitudes towards mathematics and performance outcomes revealed several salient constructs that may play a role in the underachievement of girls including: theories of intelligence, gender stereotypes, self-efficacy, mathematics anxiety, sense of belonging, and value of mathematics outside of school (Beilock et al., 2010; Good et al., 2012; Guiso, Monte, Sapienza, & Zingales, 2008; Penner & Paret, 2008). For the purpose of this paper, the term "attitudes" will refer to this list of constructs that have been linked to performance on mathematics achievement tests. This dissertation study sought to examine the relationship between these attitudes and gender differences in mathematics achievements in the context of one independent coeducational school middle school located in an upper middle-class urban neighborhood.

This chapter will discuss the results of this needs assessment. It establishes significant gender differences among students in the individual school context on all key constructs that

were measured: mathematics achievement, gender stereotypes about mathematics, self-efficacy, and sense of belonging. The results of the needs assessment shaped the focus and design of the research study and intervention. Based on the findings, two specific constructs linked to girls' underachievement in mathematics--self-efficacy and sense of belonging (Good et al., 2012; Hall & Ponton, 2005; Pajares & Miller, 1994) were chosen as the focus for a classroom-based intervention to address the gender gap in mathematics achievement.

Context

The needs assessment was conducted at the BC School, a coeducational independent school serving students in kindergarten through twelfth grade. During the year of the needs assessment 2014-2015, it had a total enrollment of 915 students; 443 boys and 472 girls. It is located in an upper-middle class urban neighborhood. The majority (about 70%) of the student population is Caucasian, and the remaining 30% identify as students of color or as having mixed racial identity. The survey respondents included students in grades 4-12. However, this study will focus on middle-school students in grades five through eight (ages 10-14) for a number of reasons. First, the limitations of the study mean that it is not possible to design an intervention for all age groups. Second, my role as the middle-school mathematics teacher provided most direct access to middle school students, making an intervention for that group most feasible. Third, students in middle school take annual standardized mathematics tests that provide valuable quantitative data for investigating possible links between beliefs about learning mathematics and mathematics achievement.

Goals and Objectives

Purpose

The primary purpose of this needs assessment was to investigate gender differences in mathematics achievement in students in grades five through eight at an independent school. A secondary purpose was to investigate the attitudes and beliefs that students in hold about learning mathematics, to determine if they differ by gender, and to see if there is a relationship between these beliefs and achievement outcomes. The key stakeholders of this needs assessment include the students in the school, the administration, teachers, and parents.

The research questions that guided this needs assessment were:

RQ1: Will middle school students' participation in supplementary single-sex mathematics classes benefit girls more than boys as measured by sense of belonging, self-efficacy and achievement?

RQ2: Will a change in girls' sense of belonging in mathematics correlate with a change in mathematics achievement test scores?

RQ3: Will a change in girls' mathematics self-efficacy correlate with a change in mathematics achievement test scores?

Methodology

Operationalization of Variables

The variables for the needs assessment included mathematics achievement as well as a number of attitudes towards mathematics that literature suggested may be linked with mathematics achievement. These included theories of intelligence, gender stereotypes, mathematics self-efficacy, mathematics anxiety, sense of belonging, value of mathematics. Mathematics achievement was measured using standardized tests results from the

Comprehensive Testing Program (CTP). All variables were measured using a 4-point Likert-like scale survey on which participants were asked to rate their level of agreement.

Table 1

Variables Used

Construct(s)	Measurement Tool(s)	Population
Mathematics achievement	CTP-4 standardized Mathematics 1 and 2 achievement test scores	Students (Grades 4-12)
Theory of intelligence	Likert-scale questionnaire	Students (Grades 4-12)
Mathematics Gender stereotypes	Likert-scale questionnaire	Students (Grades 4-12)
Mathematics self-efficacy	Likert-scale questionnaire	Students (Grades 4-12)
Mathematics anxiety	Likert-scale questionnaire	Students (Grades 4-12)
Sense of belonging in mathematics	Likert-scale questionnaire	Students (Grades 4-12)
Value of mathematics	Likert-scale questionnaire	Students (Grades 4-12)

The Comprehensive Testing Program (CTP) Tests

All middle school students in grades five through eight at the BC School take an annual battery of tests called the Comprehensive Testing Program (CTP-4) each April. The subtests include Verbal Reasoning, Vocabulary, Reading Comprehension, Writing Mechanics Quantitative Reasoning, and two sections called Mathematics 1 and Mathematics 2, which are scored as one test. All questions are multiple-choice.

The mathematics tests are based on content described in The National Council of Teachers of Mathematic (NCTM) Principles and Standards for School Mathematics (2000). This content is grouped into five main categories: (a) number sense and operations, (b) algebra, (c) geometry, (d) measurement, and (e) data analysis and probability. The Mathematics sections are primarily testing explicit information that is taught in school. The Educational Records Bureau's CTP Online Technical Report states:

> Mathematics Achievement tests assess students' conceptual understanding of mathematics, quantitative procedural knowledge, and problem solving skills. Questions that assess students' conceptual understanding typically ask students to recognize fundamental ideas in mathematics. Questions that assess students' skills in procedural knowledge tend to ask students to recall factual information about mathematics or to demonstrate understanding of basic algorithms. Questions that assess students' skills in problem solving tend to ask students to apply and integrate concepts or to identify appropriate strategies. (ERB, 2014, p. 12)

The Mathematics achievement tests are different in content from the Quantitative Reasoning subset. ERB describes the Quantitative Reasoning section as testing as student's "abilities in pattern recognition, classification, and reasoning in logic…" (ERB, 2014, p. 13). In other words, this section is less directly tied to skills that a student learning in school.

Student scores are given as both scaled scores and a percentile in different norm groups. Each subject and level has its own scale score, but unlike raw scores (the number of correct items), the "scale scores on the tests at adjacent levels of any CTP subject can be regarded as reasonably comparable, because they have been vertically equated" (ERB, 2017, p. 8). Furthermore, the scale score is an "equal interval" scale, meaning that a specific difference in

score represents approximately the same degree of difference at all regions of the scale. Since neither raw scores nor percentiles are equal interval, ERB recommends using the scale scores when looking at achievement growth over time.

In addition to scaled score, students are assigned two sets of percentile rankings, normed against a national norm group and an independent school norm group. The national norm group compares students to a representative national sample. The BC School faculty and administrators are particularly interested in the independent school norm, which compares students with other students at approximately 300 other independent schools. These norms are more competitive than the national norms, and historically students at the school in this study tend to be slightly above average in the independent school sample.

Due to the school's particular focus on the percentile ranking of the students in the independent school norm group, the needs assessment began with an investigation into the middle school students' performance based on their independent school norm rankings.[1] In 2014, the mean percentile of a middle school student[2] on the mathematics achievement test (M=59, SD=26.71) was statistically below the verbal reasoning percentile (M=65, SD=24.01) $t(192)=3.73$ $p<.01$. This prompted some concern that Mathematics achievement was lagging behind the verbal scores for middle school students. However, closer analyses of the CTP-4 scores revealed that the difference in mathematics achievement scores and verbal scores was highly gendered. The mean percentile on the independent school norms for boys on the Mathematics sub-test was the sixth percentile for boys, but 53rd percentile for girls. The pattern

[1] Subsequent analysis of CTP-4 scores was done with scale scores because these scores are an equal interval scale and are recommended for tracking achievement growth over time

[2] In this case only students in grades 5, 6, and 7 because eighth grade students take the Algebra I test

was similar for the Quantitative Reasoning subsection where the mean was 66th percentile for boys and 52nd percentile for girls. No such gender difference appeared for Verbal Reasoning. Both boys and girls had a mean verbal reasoning score at the 64th percentile.

These differences in mathematics achievement echoed the finding in the research literature that the gender gap is more pronounced among high-achieving students (Ellison & Swanson, 2010). As a benchmark, in 2014 a scale score that ranked a student at the 53rd percentile on the independent school norms corresponded to a rank in the 92nd percentile on national norms. Consequently, because the average middle school student at this school scored in the 59th percentile on the national norms, he or she would be in the top 10% of students on the national norms for mathematics achievement. This is a key point because it means that the patterns noted in the literature that appear at the very right tail of the achievement distribution such as the markedly larger gender disparity in both achievement and self-efficacy will be much more likely to be seen in this particular school setting.

Survey Design

In order to assess students' attitudes about mathematics, students were asked to complete a survey. The survey followed a Likert-style design in which participants were asked to rate their agreement with a given statement from 4 (strongly agree) to 1 (strongly disagree). The survey was a modified instrument that combined items from previously used questionnaires including The Attitude Towards Mathematics Inventory (Tapia, 1996), The Theories of Intelligence survey (Dweck, 2000), the OECD's Self-Efficacy scale (2015), and Good et al.'s (2012) Sense of Belonging Scale. It was intentionally designed with an even-numbered scale so that participants could not indicate a neutral stance on a given question. The survey was created to assess several key constructs related to mathematics attitudes including participants' theories of intelligence,

gender stereotypes about mathematics, mathematics self-efficacy, mathematics anxiety, sense of belonging in mathematics, and value of mathematics. In the course of the needs assessment, it was determined that, among these attitudes, self efficacy and sense of belonging were both more salient for understanding the relationship between gender and mathematics achievement, and potentially more affected by potential interventions in this population. Therefore, these two constructs as well as mathematics achievement are the focus of the ensuing discussion.

Questions were coded with these constructs so that inter-reliability of questions could be analyzed, but this code did not appear on the survey for participants. All questions were randomized, except for the demographic questions, which were placed at the end of the survey to avoid priming gender stereotypes. Before distributing the survey, it was reviewed multiple times by several content experts (who did not take the survey), including the Director of Research, an expert in research and data gathering, and the Director of Diversity who is trained to identify bias in testing questions, in order to discuss the instrument design, the clarity of questions, and the possibility of biases in phrasing or content. Students were asked to complete the survey electronically on SurveyMonkey on a school iPad during their mathematics class or advisory period. Students first read the consent form and signed their name electronically prior to beginning the survey. The survey was taken individually and silently to avoid discussion about the questions.

Data gathered from SurveyMonkey was exported to SPSS for analysis. Answers to the survey questions were coded as follows: *1 = strongly disagree 2=disagree, 3=agree,* and *4=strongly agree*. Gender was coded *as 1 = female* and *2 = male*. Grade level was coded as the number of the grade (e.g., *4 = fourth grade*). All survey data responses were assumed to be interval and the distance between the numbered items such as “agree” and “disagree” was

assumed to be equal with the difference between *agree* and *strongly agree*. Preliminary analyses of internal consistency reliability are high for each construct. The items in the construct sub-scales were highly correlated with the following reliability scores for students: self-efficacy subscale (α = .902), value subscale (α = .861), mathematics anxiety subscale (α = .888), theories of intelligence subscale (α = .724), sense of belonging subscale (α = .706). The overall reliability for the students questionnaire (29 items) was high with (α = .945).

Summary of Results

Research Question 1: Differences in Mathematics Achievement by Gender

Initial comparison of sub-test percentile by gender revealed a significant gender gap in mathematics. An independent t-test of means was used to compare the mean Quantitative Reasoning and Mathematics achievements of boys and girls. Although histograms of the data indicated a modest negative skew, t-tests were determined to be a robust enough test given the large sample size that results would still be reliable.

The mean achievement of girls was lower than boys at a statistically significant level on both Quantitative Reasoning, $t(192)= -3.573$, $p<.001$, and Mathematics, $t(192)=-3.398$, $p<.001$. In addition, a paired sample t-test indicated a significant difference for girls between their Verbal Reasoning scores (M=64.8, SD=22.76) and Quantitative Reasoning scores (M=52, SD=26.18), $t(105)=5.44$, $p<.001$, as well as a statistically significant difference between the girls' Verbal Reasoning scores (M=64.85, SD=22.76) and Mathematics score (M= 52.90, SD=25.46), $t(105)=5.23$, $p < .001$.

Not only the means, but also the distribution of scores was significantly different for boys and girls. The scores of both boys and girls was non-normally distributed for the verbal reasoning test with skewness of -.430 (SE=.258) and -.515 (SE=.236) respectively, indicating

moderate negative skew. For boys, Quantitative Reasoning and Mathematics scores were also negatively skewed: -.583 (SE=.258) and -.564 (*SE*=.260), respectively. Girls' scores however approximated a normal distribution closely with a skewness of -.056 (SE=.236) for Quantitative Reasoning and .063 (SE=.236) for Mathematics. These data suggested that high-ability girls, those in the upper quartile at the independent school, were underperforming compared to boys in mathematics achievement.

Histograms of CTP-4 Scores 2014

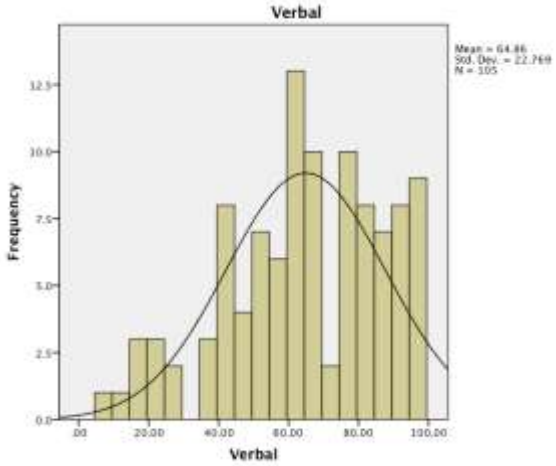


Figure 3. Verbal reasoning scores 2014 for girls.

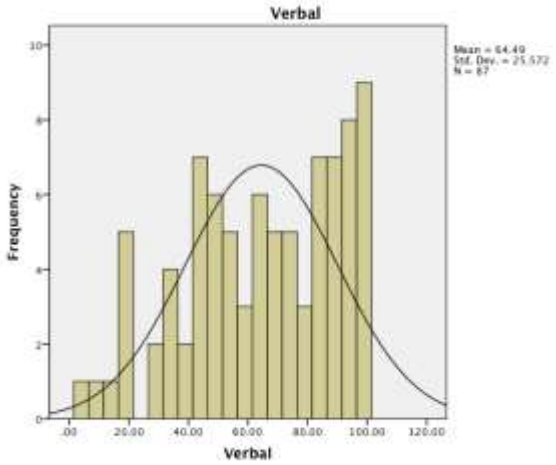


Figure 4. Verbal reasoning scores 2014 for boys.

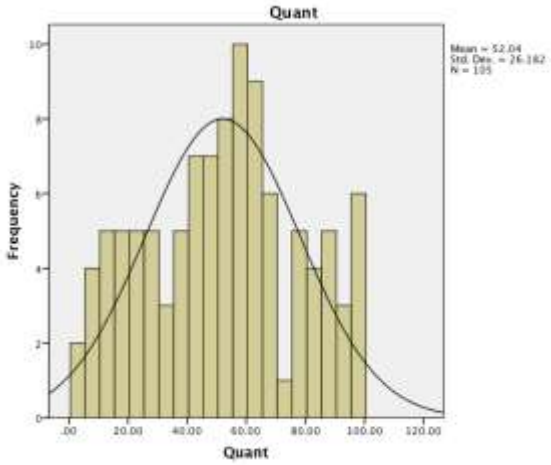


Figure 5. Quantitative reasoning scores 2014 for girls.

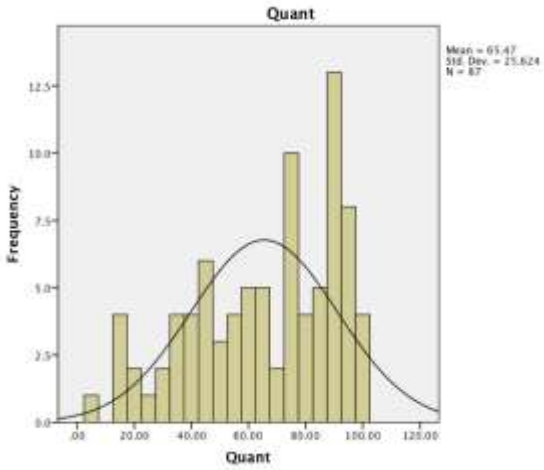


Figure 6. Quantitative reasoning scores 2014 for boys.

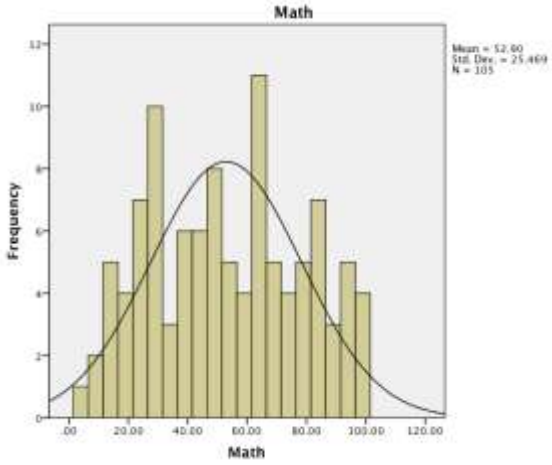


Figure 7. Mathematics achievement scores 2014 for girls.

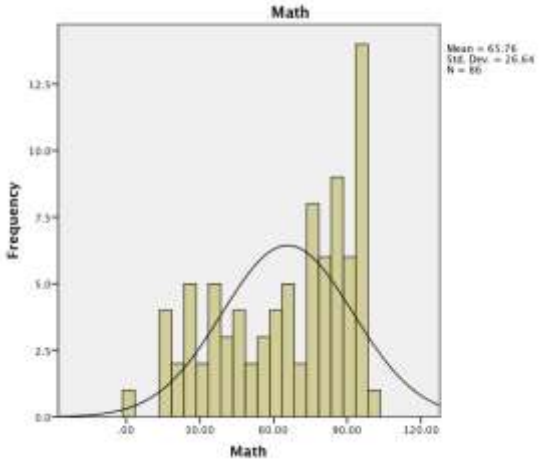


Figure 8. Mathematics achievement scores 2014 for boys.

CTP- 4 data 2015. Additional CTP-4 data were gathered in 2015 to confirm the 2014 findings using the scaled scores of students instead of the percentile scores. In 2015, eighth grade students took the Quantitative Reasoning section and were included in the analyses. As expected, the gender differences in 2015 were statistically significant for both for Mathematics ($t(198)=3.387$, $p<.01$), $d=.48$ and Quantitative Reasoning ($t(269)=5.14$, $p<.01$), $d=.63$. The effect sizes of both findings were medium to large.

Survey response data. The number of survey respondents and as well as the gender breakdown for each group of participants can be found in Table 2. Although CTP-4 standardized

test data were only available for students in grades 5-8, the survey was opened to students in grades 4-12 in order to assess if there were any notable changes in attitudes towards mathematics as students matured. This larger sample also provided greater power for statistical analyses.

Table 2

Student Survey Respondents 2015

Grade	Gender		Total
	Girls	Boys	
Fourth	18	20	38
Fifth	29	19	48
Sixth	24	24	48
Seventh	37	31	68
Eighth	18	18	36
Ninth	35	32	67
Tenth	37	33	70
Eleventh	17	25	42
Twelfth	30	14	44
Totals:	245	216	461

Research Question 2: Gender Differences on Attitudes towards Mathematics

The student survey responses revealed differences by gender on every construct subscale with varying levels of effect size. A large effect size was seen for mathematics self-efficacy, $t(461) = -6.846$, $p<.01$, $d=.64$ and mathematics anxiety, $t(461) = -5.646$, $p<.01$, $d=.54$. A moderate effect size was found for sense of belonging $t(461)= -4.165$, $p<.01$, $d=.39$, and value of mathematics, $t(461)= -4.300$, $p<.01$, $d=.39$. A small effect size was found on theories of intelligence $t(461)= -3.797$, $p<.01$, $d=.23$. Boys were also more likely to endorse traditional

gender stereotypes about mathematics, agreeing with the statement, "*In general, I think males are better at math,*" $t(461)= 3.884$, $p<.01$, $d=.36$.

Middle School Results

Subsequent analyses of student response data focused on the self-efficacy and sense of belonging of only the students in middle school (fifth-eighth grade) because I taught in this school division and the school was interested in an intervention to address the needs of students at this age. An independent t-test indicated that differences by gender in reported self-efficacy was statistically significant with a moderate effect size ($d=.48$), and gender differences in sense of belonging for middle-school students was trending towards statistically significant, with a small effect size ($d=.24$)

Table 3

2015 Results of Independent T-Test for Self-Efficacy by Gender for Fifth to Eighth Grade Students

	Gender						95% CI for Mean		
	Boys			Girls					
	M	SD	n	M	SD	n	Difference	t	df
Mean Self-Efficacy	3.05	.495	81	2.78	.626	100	-.445, -.108	-3.23	179

p=.001, d=.48

Table 4

2015 Results of Independent T-Test for Sense of Belonging by Gender for Fifth to Eighth Grade Students

	Gender						95% CI for Mean		
	Boys			Girls					
	M	SD	n	M	SD	n	Difference	t	df
Mean Sense of Belonging	3.30	.580	81	3.16	.601	100	-.316, .033	-1.60	179

p=.112, d=.24

Research Question 3: Relationship Between Self-Efficacy and Achievement

A subset of the 2015 survey data for students in fifth, sixth, and seventh graders were correlated with the students' 2015 Mathematics and Quantitative Reasoning scores. Only students in these grades take the annual standardized tests in mathematics. Eighth grade students take the Algebra I test. Mean self-efficacy scores for students in these grades had a statistically significant correlation with Mathematics scores and Quantitative Reasoning scores.

Table 5

Research Question 3 Correlations

		Mathematics scale score 2015	Mean Self-Efficacy
Mathematics scale score 2015	Pearson Correlation	1	.224**
	Sig. (2-tailed)		.003
Mean Self-Efficacy	Pearson Correlation	.224**	1
	Sig. (2-tailed)	.003	

**. Correlation is significant at the 0.01 level (2-tailed).

b. Listwise N=179

Research Question 4: Relationship between Sense of Belonging and Achievement

Mean sense of belong scores for students in grade fifth, sixth, and seventh grades also had a statistically significant correlation with Mathematics scores (r=.22, n=179, p<.01) and Quantitative reasoning (r=.20, n=179, p<.01).

Table 6

Research Question 4 Correlations

		Mathematics scale score 2015	Mean Sense of Belonging
Mathematics scale score 2015	Pearson Correlation	1	.224**
	Sig. (2-tailed)		.003
Mean Sense of Belonging	Pearson Correlation	.224**	1
	Sig. (2-tailed)	.003	

** Correlation is significant at the 0.01 level (2-tailed).
b. Listwise N=179

Conclusions

Existing data at the BC School from CTP-4 Mathematics achievement tests documented that the gender gap in mathematics achievement on standardized tests mirrors those seen nationally and internationally. The school has selective admissions, and therefore its population consists of stronger-than-average students who would tend to score in the highest percentiles nationally. The literature suggested that it is at the highest percentiles of achievement that the gender gap is in fact the largest, both in terms of girls' underachievement and in terms of self-efficacy (Ellison & Swanson, 2010; OECD, 2015; Preckel et al., 2008). Further investigation of the research literature supported the hypothesis that performance in mathematics is highly related to attitudes about mathematics including self-efficacy and sense of belonging (Dasgupta, 2011; Else-Quest et al., 2010; Good et al., 2012; Hall & Ponton, 2005; Steele & Ambady, 2006). Based on the results of this survey, there are statistically significant differences between how boys and girls experience mathematics on these constructs. As the literature would suggest, girls tended to rate themselves lower than boys on mathematics-related attitudes, and two constructs in particular—self –efficacy and sense of belonging—were correlated at a statistically significant level with mathematics achievement.

Potential Limitations

There were several limitations to this study. First, the measurement tools had limitations. The CTP-4 is a paper and pencil test that had not been revised in many years. Although the independent school norms are able to make meaningful distinctions among students at high percentiles, it is still not a highly competitive test and will not differentiate among the highest-performing students where the gender gap is likely to be the largest. Furthermore, the use of percentiles as a comparison has drawbacks because the test scores have already been transformed onto a normal distribution. Going forward, efforts were made to obtain the students' scaled scores for statistical analysis. In addition, the survey combined previously tested items from a number of questionnaires. Although it was reviewed and discussed with experts in the field, given the limited time frame it was not possible to do a full-scale pilot study before the survey was launched. The survey had four answer choices, which forced participants to choose either a positive answer (agree or strongly agree) or a negative one (disagree or strongly disagree). This decision was made to try and avoid many participants choosing a neutral position. However, the ultimately binary nature of the answer choices may distort the degree to which individuals agree or disagree with statements.

CHAPTER 3: A REVIEW OF RECENT LITERATURE

Statement of the Problem

As discussed in Chapter 1, high-ability girls are underperforming in mathematics compared to their male peers. In some regards, the gender gap in mathematics has been declining in the last fifty years (Else-Quest et al., 2010). There has been an increase in the number of mathematics and science courses that women take, and the average difference in mathematics test scores has decreased (Niederle & Vesterlund, 2010). However, while the mean effect differences between boys and girls are small (Else-Quest et al., 2010; Lindberg et al., 2010), there is still a substantial and well-documented gender gap in achievement between high ability boys and girls (Ceci et al., 2009; Ellison & Swanson, 2010). Additionally, the research literature supports a strong link between girls' lower levels of self-efficacy and sense of belonging in mathematics and lower performance on mathematics achievement tests (Else-Quest et al., 2010; OECD, 2015). Research on high-ability girls indicates that the gender gap in mathematics self-efficacy may mirror that of the achievement gap—thus widening at the right tail and disproportionately affecting girls who are high-achieving in mathematics (Hargreaves et al., 2008).

Chapter Two discussed the results of a needs assessment at an independent school. This school has a competitive admissions process and the student body is high achieving on national norms of mathematics achievement, with the average student scoring in the top 10% of students nationally. A comparison of boys' and girls' scores on the annual mathematics achievement tests revealed that boys were scoring higher compared to girls at a statistically significant level, but this same pattern was not seen with verbal achievement. Furthermore, girls reported statistically

significant lower levels of self-efficacy and sense of belonging on a self-report survey measure. The results of the needs assessment mirror the trends seen in the literature.

This chapter is focused on literature that might help elucidate potential means for closing the gender gap in mathematics achievement at the BC school. It discusses research indicating girls' mathematics achievement is reduced in a male-dominated learning environment (Inzlicht & Ben Zeev, 2000; Murphy et al., 2007) because of their heightened awareness of gender and the resulting negative effects of stereotype threat (Cherney & Campbell, 2011). This chapter then examines how an all-girls mathematics classroom could act as an intervention for improving females' mathematics self-efficacy and sense of belonging—and thus achievement—by reducing the salience of gender and negative gender stereotypes. For the purpose of this review, gender will be defined as the attitudes, feelings, and behaviors that a culture associates with a person's biological sex (American Psychological Association, 2011). This chapter will draw primarily on peer-reviewed, academic journal articles published between 2000 and 2017. Due to the limited amount of literature regarding middle school students and mathematics, this literature review will include studies involving students from elementary school to college age, as well as studies of other male- stereotyped academic subjects such as physics and economics.

Theoretical Framework

As discussed in Chapter 2, this literature review applies a social cognitivist lens to the problem of the gender gap in mathematics. Social cognitive theory is grounded on the idea that human agency, the ability to exert control over one's environment, is central to the human experience (Bandura, 2001). Research has consistently supported the hypothesis that stereotype threat reduces both students' self-efficacy and their sense of belonging (Else-Quest et al., 2010; Good et al., 2012; Steele, 1997). Negative stereotypes imply that certain groups are less welcome

or valued (Steele, 1997). Girls' exposure to stereotype threat may contribute to lower participation in mathematics activities, even among those who are the highest achievers (Good et al., 2012). Perhaps counter-intuitively, stereotype threat differentially affects girls who strongly identify with the mathematics domain (Beilock & Carr, 2005; Inzlicht & Ben-Zeev, 2000). These students are faced with a stereotype about a central element of their identity and their performance on mathematics tasks may suffer more than those who do not care as much about mathematics (Inzlicht & Ben-Zeev, 2000). Stereotype threat could provide a different explanation than the "male variability hypothesis" for the widening gender gap at the highest levels of achievements. Instead, girls who are high achieving in mathematics may be more vulnerable to negative stereotypes about women in this domain (Good et al., 2012; OECD, 2015).

Several studies bolster the theory that girls who study male-stereotyped subjects in a single-sex[3] classroom have a higher self-concept of their ability in the subject, due in part to the lower salience of gender-related self-knowledge (Eisenkopf et al., 2014; Kessels & Hannover, 2006; Preckel et al, 2008). Figure 9 shows the casual model for a single-sex classroom intervention to close the gender gap. In this model, single-sex classrooms reduce a girls' level of awareness of her gender and therefore reduces her vulnerability to the negative impacts of stereotype threat (Kessels & Hannover, 2008; Picho & Stephens, 2012). Consequently, studying mathematics in an all-girls environment may help girls to experience greater mathematics self-

[3] This research study is concerned with students' gender identity and not their biological sex, however; it will use the term "single-sex" as it is the dominant term in the literature to describe an all-boys or all-girls classroom composition.

efficacy and sense of belonging as well as higher mathematics achievement (Eisenkopf et al., 2014; Kessels & Hannover, 2008; Preckel et al, 2008; Tully & Jacobs, 2011).

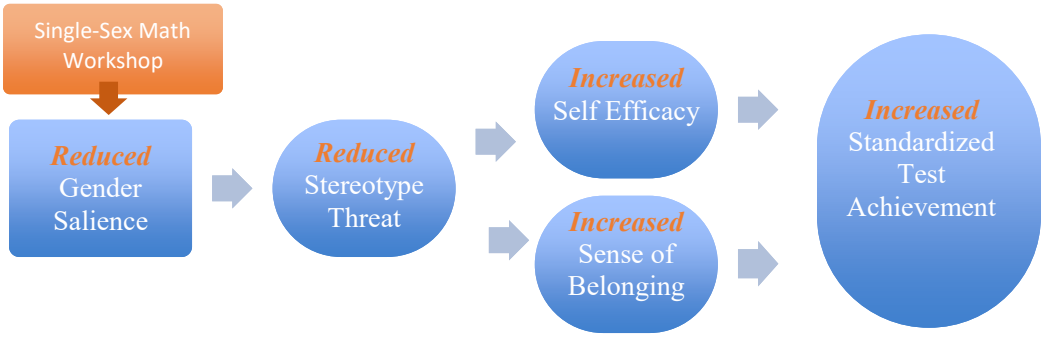


Figure 9: Intervention model.

Literature Review

Limitations of Research Methodology

Establishing a causal link between single-sex schooling and higher achievement on standardized measures has been a significant barrier to studying the potential benefits of single-sex classrooms (Mael, Alonso, Gibson, Rogers, & Smith, 2005; Pahlke et al., 2014). In the United States, enrollment in a single-sex school is voluntary; subsequently it is not possible to have a truly randomized experimental design (Signorella et al., 2013). Consequently, many research designs lack a control group for the studied population because students cannot be randomly assigned to a single-sex or coeducational setting (Cherney & Campbell, 2011; Tully & Jacobs, 2011). There is a high likelihood that self-selection of students and families to coeducational or single-sex schools contribute many confounding variables to this area of research (Park, Behrman, & Choi, 2013). Despite the fact that many studies have very large sample sizes with hundreds or even thousands of participants that lend greater credibility to their

findings (Kessels & Hannover, 2008; Preckel et al., 2008; Schneeweis & Zweimüller, 2012), results continue to be difficult to interpret without more rigorous methodological designs.

Large meta-analyses have sought to consolidate and simplify the findings in the literature on single-sex and coeducational schooling, but this may obscure important differences in how and when single-sex learning environment can be helpful. A second methodological challenge in the literature on single-sex classrooms is the large variability in dosage examined. Many studies examine differences in academic outcomes between single-sex and coeducational schools (Doris, O'Neill, & Sweetmann, 2013) while others look at the benefits of a single-sex class within a coeducational context (Kessels & Hannover, 2008). There are likely to be large differences in student school experience between having one academic course in a single-sex environment or one's entire school experience. Other research studies have examined the gender composition of a coeducational classroom (Schneeweis & Zweimüller, 2012) and suggested that there is more nuance in the gender ratio of a classroom than simple single-sex or coeducational. The literature regarding single-sex schooling has also failed to appropriately differentiate among populations. It may be that single-sex classrooms are beneficial only for certain subjects (i.e., those that are negatively stereotyped) or groups of students (i.e., those that are highly identified with a stereotyped subject).

Finally, a limitation common to studies of the relationship between mathematics self-efficacy, sense of belonging, and achievement is the reliance on self-report measures (Kessels & Hannover, 2008; Preckel et. al, 2008). Male and female students may differ in their answers to questionnaires based on socially desirable answers and stereotypes or their ability or willingness to share their beliefs (Preckel et al., 2008). This social desirability bias may be particularly true

when conducting research with middle school and high-school students who have a heightened concern regarding the opinions of their peers.

The Single-Sex vs. Coeducational School Debate

There is currently limited research on how to close the gender gap at the high end of the distribution of mathematics achievement (Ellison & Swanson, 2010; Niederle & Vestlund, 2010; OECD, 2015). One potential intervention to address the gender gap in mathematics is the creation of all-girls mathematics classes. In recent years, there has been an increased interest in single-sex schools and classrooms in public schools reflected in the creation of The National Association for Single-Sex Public Education in 2002 (Hubbard & Datnow, 2005; Williams, 2010). However, attempts to research the potential benefits of single-sex education have resulted in contradictory and inconclusive results (Tully & Jacobs, 2010).

In 2005, contracted by the US Department of Education, Mael et al. (2005) conducted a meta-analysis of 49 studies comparing the effects of single-sex and coeducational settings. Approximately one third of the studies found some benefit of single-sex education for academic achievement, while the remaining two-thirds found either no effects or mixed results. About half of the studies analyzed found some advantage for single-sex school in increasing student self-concept for both males and females, while the other half of studies found no difference (Mael et al., 2005). In response to this article, Signorella, Hayes, and Li (2013) published an article claiming that the methodology used by Mael et al. (2005) was poor and did a second analysis with different results. In this meta-analytic critique, they found no achievement or social-emotional benefits of single-sex schooling that were not correlated with "preexisting differences" such as higher socioeconomic status and cognitive performance.

There are large-scale research studies that have found no benefit of single-sex classes in mathematics for girls (Feniger, 2011) or in a few cases, better outcomes for coeducational classrooms (Doris et al., 2013; Hoffman, Badgett, & Parker, 2008). For example, Doris et al. (2013) examine the mathematics achievement of a large population of nine-year old students in Ireland (N=7,116) who attended 910 randomly selected schools. In the Irish school system, many of the local schools are single-sex schools, and parents usually elect to send their child to the local school regardless of its gender composition. Thus, the authors of this study argue that there is less of a dramatic selection effect and whether a child in Ireland attends a single-sex or coeducational school is "effectively random" (Doris et al., 2013). The researchers assessed students' mathematics achievement using a set of 25 multiple choice and short-answer mathematics questions that was tied to the mathematics curriculum.

The authors documented a significant gender gap in favor of boys, particularly at the top of the distribution; 29% of boys scored in the top quartile compared to 22% of girls. Further, their results indicated that boys who attended single-sex school performed better by 4.8 percentage points on average than boys who attended a coeducational school, but there was no corresponding benefit for girls attending single-sex school. Thus, the gender gap was actually worse between students attending single-sex schools, primarily because boys in this condition performed better. These results were maintained even when controlling for other explanatory variables included socioeconomic status and parental education. The authors did not propose any causal model for this finding. Unfortunately, because the study was not a randomized controlled study, it is not possible to rule out the possibility that there was some unknown and unmeasured difference between the students who attended the single-sex schools and coeducational schools.

Similarly it could not rule out systematic differences in the instructional approaches in those settings.

Arguments for the advantages of single-sex classes have cited both biological and sociocultural justifications (Pahlke et al., 2014; Mael et al., 2005; Signorella et al., 2013). Biological explanations are based on the claim that biologically the brains of boys and girls are different (Eliot, 2011; Halpern et al., 2007; Pahlke et al., 2014). This claim is perhaps the weakest argument for single-sex schooling because there is little evidence from neuroscience to suggest that there are reliable differences of any meaningful effect size in boys' and girls' brains that are relevant to learning (Eliot, 2011; Halpern et al., 2007). Further, there is no argument offered for why posited differences would be well served by specific and different pedagogical approaches or environments.

Other arguments in favor of single-sex education focus on sociocultural reasons. There are a number of studies that seek to investigate if single-sex schooling increases girls' achievement and academic view of themselves by offering an environment that is more supportive of girls' achievement and is protective from sexism (Pahlke, et al., 2014). In a meta-analysis of 84 of the recent studies regarding single-sex schooling, Pahlke et al. (2014) conclude, "it [single-sex classes] showed a medium advantage in middle school for girls, for both mathematics and science performance, but the effects are based on only a small number of studies and should be interpreted with caution" (p. 1065). However, on balance, Pahlke et al. (2014) conclude that there is no real advantage to single-sex classes. One particularly interesting hypothesis that these authors proposed is that the *assumptions* of the school in creating single-sex classes may be an important variable in determining potential efficacy because these messages are communicated to students. For example, a school that assumes that biological differences

between boys and girls are innate and immutable may be reinforcing gendered stereotypes about academic performance, while a school that is advocating for "girl power" by having girls collaborate in a male-dominated domain may send a different message (Pahlke et al., 2014).

Gender Composition and Stereotype Threat

Although results of studies comparing single-sex and coeducational schooling has been equivocal regarding its impact on mathematics achievement, a substantial body of experimental research conducted in the lab setting has established a link between the gender composition of a group and the activation of negative stereotypes for women regarding mathematics achievement (Inzlicht & Ben Zeev, 2000; Murphy et al., 2007). A study conducted by Inzlicht and Ben-Zeev (2000) provides striking evidence that a male-dominated gender composition of a classroom may be enough to create a threatening intellectual environment. The researchers propose that when women are outnumbered by men in a male-stereotyped activity they are more likely to experience stereotype threat and suffer a decrease in performance. In a study conducted at Brown University, 72 female undergraduates completed either a mathematics or verbal test. The tests were taken from sub-sections of the GRE. The participants took the test either in a same-sex condition (the participant and two other women in the room) or a "minority" condition (two male researchers posing as students and the female participant in the room).

An analysis of variance (ANOVA) revealed that women had a lower accuracy (number of items correct divided by number of items attempted) on the mathematics test in the minority condition (M=.55) than in the same-sex condition (M=.70). The difference in means constitutes a large effect size (Cohen's d=.80). The accuracy on the tests was controlled for previous achievement using self-reported SAT scores. The same disparity between scores was not found

for women taking the verbal test. These research findings how a powerful implication; woman may experience a threatening learning environment when they are outnumbered by men, even if there is no explicit mention of a negative stereotype.

In a follow-up experiment, the authors had 92 male and female undergraduates take a GRE mathematics section in one of three settings: a single-sex female group, a mixed-gender group in which females were the minority, or a mixed-gender group in which females were in the majority. As the authors predicted, female participants in the minority group scored the lowest (M=.58), female participants in the majority group scored slightly better (M=.63), and female participants in the single-sex group scored the highest (M=.70). Females' mathematics achievement decreased as the number of males present increased. There was no relationship between gender composition and male scores.

This second experiment adds important nuance to the potential conclusions. Based on these findings, stereotype threat is not either "on" or "off" based on the gender composition of the environment. It may have a linear relationship in which the degree of stereotype threat increases as the ratio of males to females increases. Furthermore, this could help make sense of the widening achievement gap on high-stakes competitions on which girls are generally much outnumbered compared to boys. Despite the powerful implications of the study, there are several important limitations to consider. The sample size of both studies was relatively small and the sample was taken from a group of self-selected high-achieving college-aged students. It is not possible to assume these results would be the same with a more diverse population or with much younger students. However, this is also consistent with findings that suggest sociocultural factors contributing to the gender gap are highly salient for very high-achieving girls.

One potential way to reduce the impact of these gender stereotypes and limit stereotype threat for females is through changing the gender composition of specific, targeted, learning environments without creating a completely single-sex environment. One study by Austrian behavioral economists Schneeweis and Zweimüller (2012) suggested that girls educated in environments with a higher proportion of other girls may be less susceptible to detrimental effects of negative gender stereotypes later in life. The study examined the relationship between the school choice that girls (n=7,472) made for high school and the gender composition of their elementary school (grades 5-8). The authors used data from 19 cohorts of students from Linz, Austria. The results indicated that girls were more likely to choose a male-dominated school type (technical school) if their elementary school classes had a larger share of females. Given the very large sample size of the study, the results may indicate that a classroom need not be entirely single-sex in order to help females combat gender stereotypes. The results support the hypothesis that simply having a higher proportion of females in a classroom may help reduce the salience of gender as an identifying variable.

Gender Composition and Self-Efficacy

Research suggests that the gender composition of learning environments can have an impact on girls' achievement precisely by impacting their self-efficacy. A number of research studies support the theory that girls who study male-stereotyped subjects in a single-sex classroom have a higher self-concept of their ability in the subject, due in part to the lower salience of gender-related self-knowledge (Cherney & Campbell, 2012; Eisenkopf et al., 2014; Kessels & Hannover, 2006; Preckel et al., 2008).

In a classroom-based study, Cherney and Campbell (2011) examine the effect of stereotype threat and the role of gender composition in a classroom on mathematics achievement.

The study involved boys (n=209) and girls (n=339) from coeducational and single-sex high schools. All of the students from single-sex school attended private institutions and half of the students at coeducational schools attended private institutions. The participants completed a 10-question content-knowledge mathematics test designed by three high-school teachers as well as the Achievement Motivation Scale and the Rosenberg Self-Esteem Scale. Results were analyzed with multivariate analysis of variation (MANOVA) and indicated that girls in the single-sex schools had statistically higher self-esteem and achievement motivation than the females in coeducational schools. For both genders, performance on the mathematics test was higher for students in the single-sex classes. The authors propose that implicit and explicit cues of negative stereotypes are minimized in a single-sex classroom. Given that students were not randomly assigned to classrooms, this study cannot rule out the possibility that differences were due to confounding factors that may play a role in families selecting single-sex or coeducational settings (Cherney & Campbell, 2011). Despite this limitation, the study used a large sample size and provided some suggestive evidence that single-sex classrooms for mathematics may positively affect self-efficacy in mathematics for female students.

In a quasi-experimental design with physics students, Kessels and Hannover (2008), educational psychologists at a research university in Berlin, conducted a study in which they randomly assigned eighth grade (n=401) students in a coeducational school to take physics in either a mixed or single-sex classroom. The authors argued that both physics and mathematics are stereotyped as masculine, and they wished to investigate if girls' subjective self-concept in physics would benefit from an all-girls learning environment. A Likert-style questionnaire of self-concept for physics ability was administered. Results were analyzed using factorial analysis of variance and indicated that girls in the single-sex class reported a statistically significant

higher self-concept of ability than girls in the coeducational class. The effect size of this difference was moderate (d=.48). Boys had a higher self-concept of ability compared to females regardless of which class they were assigned. The authors proposed that these results could be due to the lower awareness of gender as a distinguishing characteristic in the single-sex class. In order to test this hypothesis, the authors used a computer-based measurement tool that presented an adjective on the screen. Participants pressed a key to indicate *yes* or *no* to indicate if the word described them. Boys and girls both tended to endorse more gender-stereotyped traits in mixed-gender classes than single-sex classes. As expected, girls in single-sex classrooms had a slower response time for judging traits as feminine than girls in coeducational classes.

Although this study was conducted with students in physics and not mathematics, the authors propose that the same effect would be present for any subject stereotyped as masculine, including mathematics. The results of the study provide a possible explanation for improved achievement outcomes for female students in single-sex mathematics class. Essentially, gender is not an important distinguishing characteristic in an all-females classroom and therefore gender-related traits are less readily called to mind (Kessels & Hannover, 2008). This study is particularly important because it is one of the few examples of research on the implementation of single-sex classes within a coeducation environment.

A second school-based study was conducted by Eisenkopf et al. (2014) in a Swiss high school. The authors examined the role of single-sex classes on mathematics grades and self-efficacy. The study compared girls' school mathematics grades in single-sex (n=122) and coeducational (n=375) mathematics classes within the same Swiss high school. The school had a disproportionate number of girls enrolled, and decided to create some all-girl classes in order to try and balance the gender composition within the coeducational classrooms. A comparison of

means indicated that girls in the single-sex classes performed better at a statistically significant level than students in the coeducational class. Girls in the single-sex mathematics class had an average grade of 4.67 (out of 6) compared to 4.50 for girls in the coeducational class. The same effect was not seen for German class, a subject area that the authors use as a control because it may not be as vulnerable to gender stereotypes. Importantly, the effect size of single-sex classes was greater for females who were high performing in mathematics. These results support the fact that girls in high-ability groups might benefit even more from all-girls mathematics classrooms than average-ability students (Ellison & Swanson, 2011). In a second phase of the experiment, 213 girls, 62 in single-sex classes and 151 in coeducational classes, completed a survey that included questions about self-efficacy and self-assessment of competency in mathematics. Those students in the single-sex classes rated both their ability in mathematics and their sense of self-concept in mathematics as higher than those in coeducational classes. There was no statistical difference when the same survey was given to students regarding German class. The comparison to the German class as a control group is compelling evidence that an all-female class is beneficial for females in male-stereotyped subjects such as mathematics.

Research has also sought to investigate if stereotype threat functions similarly in non-western cultures and in different school contexts. Picho and Stephens (2012) conducted a quasi-experimental study in which Ugandan high-school girls (n=89) were randomly assigned to one of two conditions—one where they took a mathematics achievement exam under conditions designed to prime negative stereotypes or a second, neutral control condition. In the stereotype primed condition, the students were told that the mathematics test had shown differences in performance between boys and girls; in the control condition the girls were told only about the test format. The study participants also answered questions regarding their level of self-

identification and self-efficacy in mathematics. Half of the participants attended a single-sex school and half attended a coeducational school.

First, the authors conducted several t-test to determine if there were differences in mathematics identification and self-efficacy between the two schools. They found that girls who attended the single-sex school reported statistically significant higher levels of self-identification and self-efficacy in mathematics and also performed better on the mathematics achievement test. Next, the authors used an analysis of variance (ANOVA) to test if stereotype threat would affect the performance of the students in the two difference schools and conditions. They documented that the difference between the treatment and control conditions was statistically significant at the coeducational school but was not statistically significant at the single-sex school. In other words, students who took the mathematics achievement test under heightened stereotype threat at the single-sex school did not do significantly worse on the achievement test than those who did not have stereotype threat primed. Conversely, at t-test of independent means indicated that girls at the coeducational school did considerably worse in the stereotype threat condition than in the control condition. The authors propose that stereotype threat only operated in the coeducational school setting because negative gender stereotypes are more accessible in coeducational contexts. The methodology of the research contributes to the literature by comparing stereotype threat and control conditions in two different school contexts. One limitation of this study is that it did not seek to control for other contextual factors that could also explain the differences in performance outcomes between the two schools. Even so, the results of this research study reinforce the hypothesis of Kessels and Hannover (2008) that gender stereotypes may be more consciously accessible in a coeducational school environment.

Qualitative research methods have also added important findings in the research on girls' self-efficacy and mathematics. Tully and Jacobs (2011) investigated self-efficacy in mathematics among university students. They used a mixed-methods approach to examine the role that an all-females high-school mathematics experience may have had on students' subsequent self-perceptions of their mathematics ability, and on the likelihood of these students pursuing engineering at university. The study was conducted at an Australian university with 112 students (39 females, 73 males), who were undergraduate engineering majors. The authors conducted structured and semi-structured interviews. Seven out of ten women interviewed indicated that high self-efficacy in mathematics was a key reason they chose engineering as a major. Male students interviewed did not mention being good at mathematics as a top reason for choosing engineering. In addition, of all the students who participated in the study, women who attended a single-sex high school demonstrated the highest self-assessment of mathematics ability, higher compared to all participating boys. Although the sample for this study was not large, the qualitative element of the research suggests some concrete reasons why single-sex classes might benefit female mathematics students. Girls reported feeling empowered by a single-sex classroom and reported a much higher level of mathematics self-efficacy potentially attributable to having been in that kind of environment.

Gender Composition and Sense of Belonging

There is growing evidence that the gender composition of an environment may be a causal factor in determining if the negative stereotype of women's weaker abilities in mathematics will be activated and a students' sense of belonging threatened (Inzlicht & Ben-Zeev, 2000). In one study, Murphy et al. (2007) found that undergraduate women were less likely to express interest in attending a conference when it was portrayed as having a male

dominated (3:1) ratio of participants than when it was advertised as having equal numbers of male and female attendees. The authors conducted the experiment with 25 males and 22 female undergraduate students, all of whom were majors in mathematics, science, or engineering and self-identified as being good at mathematics tasks. The participants viewed a 7-minute video about a potential conference. One version of the video showed approximately 150 people with a ratio of 3 men for every 1 woman. A second version depicted the conference as having equal men and women.

The authors recorded cognitive and physiological responses to the video. Analysis of variance (ANOVA) indicated that female participants who watched the video in which women were outnumbered showed higher heart rate and skin conductance. Sense of belonging was measured using a Likert scale questionnaire that asked the participants to rate their anticipated sense of belonging at such a conference. Female participants reported less interest in participating in the conference where males were the majority and a lower sense of belonging, at a statistically significant level. The gender representation in the video did not have any effect on the men's anticipated sense of belonging. The findings of the study lend support for the hypothesis that women experience stereotype threat in gender-unbalanced settings and experience a lower sense of belonging. Furthermore, the results suggest that identity threat can be experienced, or even potentially experienced more, by women who have high confidence in mathematics. Although the same size for this study was not large, it helps to explore the potential ways in stereotype threat could be cued in a situation by underrepresentation of women.

Another study investigating gender composition of a field and women's sense of belonging found that women tend to perceive that they expend more energy in STEM fields than men do. Smith et al. (2012) conducted a study with graduate students in STEM fields (n=149),

and asked them each to fill out a survey regarding their level of effort expended in their studies compared to their peers and their sense of belonging in their field. Female students reported that they expended more effort than their male peers, and this level of effort was positively correlated with their self-reported sense of belonging. Male students' perceived level of effort expended was not correlated with sense of belonging. In a follow-up study, the authors found that when they advertised a fictional graduate program in "eco-psychology" as male dominated by using a brochure with majority men pictured compared to a more gender-balanced brochure, women reported higher levels of expected effort expenditure and lower interest in the program. The authors conclude that women interpret a numerical underrepresentation of women to suggest that they will have to work harder than their peers to succeed in the domain; and that women in these circumstances are less likely to pursue this field of work. This research study makes an important connection between gender imbalances in professional settings and the underrepresentation of women in STEM careers.

Conclusion

This chapter discussed literature supporting the hypothesis that the gender composition of a learning environment affects girls' self-efficacy and sense of belonging in subjects that are stereotyped as male-dominated, including mathematics (Inzlicht & Ben-Zeev, 2003; Murphy et al., 2007). Then, research was reviewed that indicated single-sex classes in a male-dominated domain may help girls to experience greater mathematics self-efficacy and sense of belonging as well as higher mathematics achievement (Eisenkopf et al., 2014; Kessels & Hannover, 2008; Preckel et al, 2008; Tully & Jacobs, 2011). One explanation for this finding is that single-sex classrooms reduce a girls' level of awareness of her gender and therefore reduces her

vulnerability to the impact of stereotype threat (Kessels & Hannover, 2008; Picho & Stephens, 2012).

Chapter Four will discuss an intervention at The BC School to address the gender gap in mathematics achievement. Drawing on the research literature, the school piloted a single-sex mathematics course once during each eight-day cycle of classes in 2015 to 2016, for 20 50-minute sessions between September and June. This "math workshop" was an additional ungraded and mixed-ability mathematics class that was added to the students' schedules. Regular mathematics courses continued to meet daily in coeducational, ability-grouped sections. The pilot year was completed with fifth and seventh grade students, and the program was repeated with some modifications for fifth and seventh grade students in 2016 to 2017.

CHAPTER 4: INTERVENTION PROCEDURE AND PROGRAM METHODOLOGY

Introduction

As discussed in Chapter 3, educational researchers have conducted numerous studies on the potential benefits of single-sex classes for girls in masculine stereotyped subject matter including mathematics (Cherney & Campbell, 2011; Eisenkopf et al., 2014; Schneeweis & Zweimüller, 2012). However, much of the prior research on the effects of stereotype threat have either been in a laboratory setting (Inzlicht & Ben-Zeev), or in the cases of applied research, have explored the potential benefits of single-sex schooling by comparing schools that are completely coeducational or single-sex (Cherney & Campbell, 2011; Doris et al., 2013; Picho & Stephens, 2012). Studies on all-female classes for male-stereotype subjects in a coeducational school indicate that this type of intervention might benefit high-ability girls who strongly identify with mathematics (Eisenkopf et al., 2008; Kessels & Hannover, 2008). This chapter describes an intervention at the BC School, an independent school in an upper-middle class urban neighborhood, which provided single-sex mathematics groupings once during each eight-day cycle of classes in addition to daily coeducational mathematics classes.

Purpose and Research Questions

The purpose of this study was to determine if the addition of twenty 50-minute single-sex middle school mathematics classes to the regular mathematics instruction would help to close the gender gap in mathematics self-efficacy, sense of belonging and achievement. This study addressed three primary research questions:

RQ1: Will middle school students' participation in supplementary single-sex mathematics classes benefit girls more than boys as measured by sense of belonging, self-efficacy, and achievement?

RQ2: Will a change in girls' mathematics self-efficacy correlate with a change in mathematics achievement test scores?

RQ3: Will a change in girls' sense of belonging in mathematics correlate with a change in mathematics achievement test scores?

This chapter discusses the intervention procedure, design, and methodology. It begins with information about the research setting and participants. Next, it describes each of the data sources that were used to measures students' self-efficacy, sense of belonging, and mathematics achievement. This chapter then moves into a detailed description of both the pilot year and intervention year procedure as well as a detailed description of how the data were collected and analyzed. Finally, the chapter concludes with a discussion of the strengths and weaknesses of this intervention design as well as the project effect size. A timeline of the intervention from the needs assessment through data analysis is included in Table 7.

Table 7

Timeline of Research Activities

Activity	Date Completed
Needs Assessment: Survey of attitudes towards mathematics and analysis of 2014 and 2015 CTP-4 scores	April 2015
Math workshop intervention pilot year	September 2015-June 2016
Pilot data collection: Survey of attitudes towards mathematics and analysis of 2016 CTP-4 data	April-June 2016
Begin math workshop intervention	September 2016
Pre-intervention survey of students' mathematics self-efficacy and sense of belonging	September 2016
Classroom observations using RTOP	January 2017
Student interviews	April-June 2017
CTP-4 Testing and scores collected	April 2017
Post-intervention survey of students' mathematics self-efficacy and sense of belonging	May 2017
Intervention data analysis	May-July 2017

Participants and Setting

The research was conducted at the BC School, a coeducational independent school, serving students in pre-K through twelfth grade. Total enrollment in 2015-2016 was approximately 920 students, and then rose slightly in 2016- 2017 to 960 students. The school is located in an affluent urban neighborhood in a metropolitan area. The majority of students are white (70%), with approximately 30% identifying as students of color. Students in fifth and seventh grades participated in the intervention. In 2015 to 2016,during the pilot year, there were 65 students in the fifth grade and 75 students in the seventh grade. For the intervention year 2016 to 2017, there were 65 students in fifth grade and 71 students in seventh grade (Table 8). All

students in the fifth and seventh grades participated in the single-sex mathematics workshop ("math workshop"). There were four sections of math workshop in fifth and seventh grade: two sections of girls and two sections of boys. The class size ranged from 14-19 students in a section.

Middle school students at the BC School have mathematics class every day for 50 minutes during the school's 8-day cycle. Mathematics is the only subject that is "tracked" or grouped by prior achievement, and fifth grade is the first grade in which this practice of homogeneous groupings is begun within the BC school. In fifth grade, there are two levels of mathematics, on-level fifth grade, and advanced fifth grade. In seventh grade, students are further divided into three levels: pre-algebra, advanced pre-algebra and Algebra I. In contrast, the math workshop classes met in heterogeneous-ability groups with students from all different levels.

Table 8

Number of Participants in Pilot and Intervention Years: Pilot Year 2015-2016 Participants

	Boys	Girls	Total
Fifth Grade	37	28	65
Sixth Grade	36	34	70
Seventh Grade	38	37	75
Eighth Grade	34	39	73

Table 9

Intervention Year 2016 to 2017 Participants

	Boys	Girls	Total
Fifth Grade	32	33	65
Sixth Grade	38	37	75
Seventh Grade	35	36	71
Eighth Grade	35	38	73

The Researcher's Role

As the primary investigator, and the middle school mathematics specialist, I was both a researcher and a participant in the study. I influenced the study design and findings, and served as the math workshop instructor for female students in fifth and seventh grades. At times, the participant-researcher role can create tension between the researcher and participants. My understanding of this conflict was informed by the work of Glesne and Peshkin (2010). For example, the observations may cause participants to feel as if they are being spied upon, or they may resent being part of an experiment. However, my preexisting role in the community and my ongoing work with students lessened this tension. Another challenge for the participant researcher is managing bias when conducting data analysis, especially for qualitative data such as document analysis and open-ended interviews (Chenail, 2011). In this case, having a pilot study provided a means for me to test the methods and see if the procedures function as planned and to address concerns regarding researcher bias (Chenail, 2011). For this research project, feedback, and data-collection from the pilot year version of the mathematics workshop informed the design and implementation of the intervention. For each data source, this chapter will clarify any differences between the pilot year and intervention year methodologies.

Data Sources

CTP-4 Standardized Mathematics Tests

All middle school students at the independent school take an annual battery of tests in April called the Comprehensive Testing Program (CTP-4). The subtests include quantitative reasoning, and two sections of mathematics, "Mathematics 1 and 2," which are scored as one section. The Quantitative Reasoning section is intended to focus on conceptual knowledge, while the two mathematics sections are aligned with content specific-skills. The mathematics content is

aligned with the Common Core State Standards in Mathematics (CCSSM) (Clune, 2014). Test results are reported in a number of formats. First, students are given vertical scaled scores for each sub-test that allows for tracking individual progress across years. Second, a student's scores are normed and reported as percentiles compared to other students nationally (national norm) and to other students at independent schools (independent norm). The school leadership is primarily interested in tracking student progress as annual growth on the scaled score measure and as independent school percentiles. Achievement scores from the online version of the CTP-4 Mathematics 1&2 achievement test were used in both the pilot year and intervention year evaluation.

Modified Mathematics Attitude Survey

Pilot year. In order to allow comparison between student responses with data from the needs assessment, in the pilot year the same survey instrument was used with students who received the math workshop program (Appendix A). As discussed in Chapter 2, the survey used during the pilot year was a modified instrument that was used in the need assessment and combined items from previously designed questionnaires including The Attitude Towards Mathematics Inventory (Tapia, 1996), The Theories of Intelligence survey (Dweck, 2000) and the OECD's Self-Efficacy scale (2015) and Good et al.'s (2012) Sense of Belonging Scale. It was a relatively long survey with twenty-nine statements. As with the needs assessment, all statements were randomized except for demographic questions, which were placed at the end.

Intervention year. In response to the pilot year data collection, the survey was refined to focus more specifically on students' self-efficacy and sense of belonging and to be shorter and easier for students to complete (Appendix B). Several teachers during the pilot year mentioned that there were terms or phrases that fifth grade students did not understand. There were also

several items that were identified as “double-barreled” (i.e., I feel uncomfortable and out of place in math class) which were edited to include only one statement for clarity. The new tool used items from The Attitude Towards Mathematics Inventory (Tapia, 1996), the Fennema-Sherman Mathematics Attitude Scale (1976) and Good et al.’s Sense of Belonging Scale (2012). It had a total of 12 statements regarding mathematics and 2 demographic questions about gender and grade level.

Classroom Observations

Pilot year. During the pilot year, the school psychologist observed the math workshop sessions several times informally. However, there was no formal observation tool used. In order to have a more reliable means for collecting data about the teacher behavior and classroom culture, an observation tool, The Reformed Teaching Observation Protocol (RTOP) was selected for the intervention year.

Intervention year. Classroom observations were conducted using the Reformed Teaching Observation Protocol (RTOP) (Appendix C). The boys’ and girls’ sections of mathematics workshop met simultaneously, therefore it was not possible to have the same teacher for both groups. The goal of the RTOP observations was to gather information about similarities and differences in the two teachers’ classroom behavior that could play a role in the student experience and outcomes. It is a 25-item classroom observation protocol that that uses a five point scale (0-4) to assess the degree to which classroom instruction demonstrates standards-based, student-centered and inquiry-oriented practices (Sawada et al., 2002). It was designed to be used for classroom of any age group from Kindergarten through college. For each statement, the observer must give a score from 0 (never occurred) to 4 (very descriptive). Internal consistency measured with Cronbach’s alpha has been high, with correlation coefficient in the

range of 0.88 to 0.97 between different rater's RTOP scores for classrooms (Sawada et al., 2002). The RTOP consists of three sections; five statements regarding lesson plan and implementation, 10 statements regarding the content of the lesson, and 10 questions regarding the classroom culture.

Three school administrators were trained in the use of RTOP. First, a middle school mathematics teacher was video-recorded teaching; administrators were given one week to independently watch the video and score the lesson using RTOP. Next, the scores were reviewed in a group discussion with me to discuss scoring discrepancies and clarify observers understanding of the descriptions for each item. For example, a statement such as "the lesson was designed to engage students as members of a learning community" is purposively vague. As a group, the school administrators and I agreed upon what types of behaviors/activities would indicate that this statement was descriptive of the classroom. Following this, the administrators viewed a new video lesson of the same middle-school mathematics teacher and again scored the lesson independently with the RTOP scale. I reviewed the scores to assess the reliability of the judgments made by the observers. Next, I calculated the intraclass correlation coefficient (ICC; using a two way random effect models) among the three trained observers. The ICC was .77; therefore, after discussion with the observers regarding remaining discrepancies in the scoring, classroom observations were scheduled (table 4.1).

The three trained school administrators observed the two math workshop teachers multiple times during January 2017. Due to the last-minute cancellation of an observation, one teacher was observed three times and the other four times. Both teachers were observed in fifth grade and seventh grade all boy or all girl classes. The completed RTOP ratings were returned to

me, and these results were not analyzed until the completion of the intervention. The results of these observations will be included in the data analysis section of Chapter 6.

Participant Interviews

Pilot year. During the pilot year, I conducted several semi-structured interviews with fifth and seventh grade girls who had participated in the pilot (Appendix D). A semi-structured interview involves a set of predetermined questions that are established prior to the interview but which may be added to or replaced in the course of the interview (Glesne and Peshkin, 2010). I selected these students by assigning each girl in the pilot study a number and then using a random number generator to select twelve students, three from each of the four girls' sections (2 in fifth grade and 2 in seventh grade) of math workshop. All students and their parents signed informed consent to participate and to be recorded. The interviews were conducted during a time of the student's choosing such as a break or immediately after school ended for the day. The interview was recorded on an iPad using the application Dictate and Connect. The majority of interviews lasted approximately 15 minutes, although a couple went as long as 45 minutes.

Through listening to the audio recordings of the interviews, there was evidence that I had difficulty adhering to best practices for conducting research interviews. Ideally, an interviewer will try to remain as neutral as possible and will keep the conversation from straying too far off-topic (Turner, 2010). The dynamic between the interviewees and me was uncomfortable at times because I was also the students' math workshop teacher and students may have been reluctant speak openly with me about their views. In this participant-researcher situation, my lack of experience conducting interviews coupled with my closeness to the study population introduced a problem of researcher bias (Chernail, 2011). Therefore, the results from these interviews were not included in the pilot year evaluation. Instead, they were only used to help improve the design

of the intervention year interview process. The questions from the pilot year interviews were revised for the intervention year to better align with the research questions, and a school administrator with research experience was asked to conduct the interviews for the following year to reduce bias and to encourage students to feel more comfortable sharing both positive and negative feedback.

Intervention year. During the intervention year, open-ended interviews were conducted with six fifth grade students and six seventh grade students, three from each of two girls' math workshop sections. The open-ended interviews were highly structured, with each participant asked the exact same question with the same wording (Appendix E). However, questions were phrased such that respondents can give as much detail as they would like (Turner, 2010). In order to reduce researcher bias, the Director of Research, who has a doctoral degree in an education-related field and was experienced conducting research interviews, conducted all of the interviews with students.

To select the students to be interviewed, each student was assigned a number and the Director of Research used a random number generator to choose three girls from each of the four sections of math workshop. Students were then informed with a written note from the Director of Research that they were invited to participate in an interview, but that participation was completely voluntary. They were also asked to bring home a permission slip that both they and their parents or guardians would sign. One student chose not to participate and a new student name was randomly selected. All of the other students agreed to participate and received parental permission. The identity of the student that had been selected was not shared with me until May of the academic year in order to reduce any tendency to alter behavior in class towards these students.

In order to standardize the interview experience the interviews all took place during study hall at the end of the school day between 3:10 and 3:40. All interviews were approximately 10 minutes long. They were conducted in the Director of Research's office, which provided an environment of confidentiality and minimal distraction There were no other students present in that office. The office was shared with one other faculty member who was occasionally present for the interviews, but who was not familiar with the students or the math workshop program. The students were each asked the identical five questions: two regarding their self-efficacy in mathematics, two regarding their sense of belonging in mathematics and one reflecting on differences between their experience in math workshop and their "regular" daily mathematics class.

The Director of Research did not engage in a back-and-forth conversation with the students and instead used occasional neutral phrases such as "tell me more about that" or "can you give me an example?" to encourage a student to elaborate on a point. The goal of her questions was to keep the interviewee on topic and to obtain as thorough a response as possible (Creswell & Clark, 2007).

Table 10

Summary Matrix of Data Collection and Analysis

Question	Method	Data	Analysis
RQ1 Will middle school students' participation in supplementary single-sex mathematics classes benefit girls more than boys as measured by sense of belonging, self-efficacy and achievement?	Achievement scores, survey responses, interviews	CTP-4 scores, survey responses, interview transcripts	T-test of independent means, thematic analysis
RQ2 Will a change in girls' self-efficacy correlate with a change in mathematics achievement test scores?	Achievement scores, survey responses	CTP-4 scores, self-efficacy survey responses	Correlation analysis
RQ3 Will a change in girls' sense of belonging correlate with a change in mathematics achievement test scores?	Achievement scores, survey responses	CTP-4 scores, sense of belonging survey responses	Correlation analysis

Procedure

Pilot Year

During the pilot year, the math workshop lessons were targeted towards specific skills that the middle school mathematics department chair identified as weaknesses for students in the middle school based on the results of the CTP-4 mathematics tests (Table 11). These outlines were created without consultation with other mathematics teachers or with me. The department chair viewed the course as an opportunity to practice skills tested on the CTP-4 as well as test-taking strategies. He was also the instructor for the boys' seventh grade math workshop, while

another male mathematics teacher taught the boys' fifth grade section. In order to differentiate the materials among the different prior ability levels, the mathematics department chair prepared multiple worksheets of varying challenge levels. He was not interested in using math workshop as an opportunity for more open-ended problems or for situating mathematics in a more real-world context. Much of the materials for the seventh grade pilot were original problems written by the mathematics department chair and consisted of lengthy sets of computation. I worked more collaboratively on materials for the fifth grade math workshop with the boys' instructor. These materials were still skill-focused, but tended to emphasize more collaboration among students and application to real-life scenarios.

Table 11

The Outline of Topics for the Pilot Year Math Workshop Program: Math Workshop Grade 5, 2015-2016

Topic	Duration (cycles)
Rounding & Estimating	2
Powers of 10 & Metric Conversion	2
MIXED PROBLEM SETS	1
Factors and Multiples	1
Ratios and Fractions	2
Percentage	1
MIXED PROBLEM SETS	1
Probability	2
Algebraic Representation	1
Geometry	1
Substitution, and other Multiple Choice Tactics	2+

Table 12

The Outline of Topics for the Pilot Year Math Workshop Program: Math Workshop Grade 7, 2015-2016

Topic	Duration (cycles)
Rounding & Estimating	2
Scientific Notation	1
Metric Conversion	1
MIXED PROBLEM SETS	1
Factors and Multiples	1
Ratios and Fractions	2
Percentage	1
MIXED PROBLEM SETS	1
Probability	2
Algebraic Representation	1
Geometry	1
Substitution, and other Multiple Choice Tactics	2+

Intervention Year

Although still based on an analysis of student CTP-4 data, the curriculum and pedagogy of the math workshop program was significantly different during the second year. The lessons were designed to be more collaborative and student-focused. In July and August 2016, the two mathematics workshop instructors and a new head of the middle-school mathematics department, met to review the results of the April 2016 CTP-4 scores and to look for patterns in content mastery. The instructors identified topics in the curriculum that seemed to be relative weaknesses and that could potentially benefit from additional instruction. For example, probability was determined to be a weaker content area for a majority of students, and therefore several sessions of mathematics workshop integrated key ideas in probability. The first 16 mathematics workshop classes occurred before the annual CTP-4 testing, and they aimed to address these specific concepts and skills on which students' might need more exposure and practice. After the CTP-4

testing, the remaining four classes in each grade were devoted to an end-of year project (Table 13). Fifth grade students worked in small groups to solve a riddle and then to explain their solution both in a diagram or written format as well as with an oral presentation. Seventh grade students worked in small groups to construct a scale model of the classroom.

Table 13

Outline of Math Workshop Lesson Topics: Math Workshop Grade 5, 2015-2016

Topic	Duration (cycles)
Estimation	2
Measurement	2
Powers of Ten	1
Percentage	2
2D Figures: Symmetry, Rotation	2
3D Figures: Surface Area, Volume	2
Pre-Algebra	2
Statistics: Reading Graphs	1
Probability	2
Project: Riddles	4

Table 14

Outline of Math Workshop Lesson Topics: Math Workshop Grade 7, 2015-2016

Topic	Duration (cycles)
Estimation	2
Measurement	2
Geometry	4
Probability	4
Algebraic Patterns	4
Project: Scale Model of Classroom	4

The two math workshop teachers worked together to design the lessons during five paid curriculum development days in July and August 2016 so that the boys' section and girls' section used the same materials, covered the same topics, and took part in the same activities. All lesson plans and materials were shared via Google Drive. A male teacher taught the boys in fifth and

seventh grades and I, a female teacher, taught the girls in fifth and seventh grade. Unlike the students' daily mathematics classes, which were divided by mathematics ability level, the mathematics workshop was heterogeneous regarding ability. Scheduling constraints necessitated that multiple "regular" mathematics sections be grouped together during mathematics workshop. The workshop's materials were highly differentiated with multiple levels of difficulty so that students at a variety of skill levels could learn new material that built on previous knowledge. Activities were chosen that met the description of "low floor, high ceiling," meaning that students with less prior knowledge of a subject were able to find a means for making sense of the problem, but students with more prior knowledge could also apply more advanced, sophisticated methods of mathematics as well.

The instructional practices of both sections were informed by research that suggests some important, tangible practices that may reduce stereotype threat and improve mathematics self-efficacy. These include positive verbal feedback, collaboration with other students, and the presence of role models (Ellison & Swanson, 2010; Marx & Roman, 2002; Preckel et al., 2008). Preckel et al. (2008) suggest that focusing on providing positive feedback and creating a supportive environment of teachers and peers is a potential strategy for intervention. Students worked exclusively in pairs or small groups and receive positive, verbal encouragement during class discussion and in one-on-one interactions. The classes did not include any competitive activities. Research by Niederle and Vesterlund (2010) found that male students tend to be more interested in competition and to outperform female students in coeducational competitive circumstances. They further propose that highly competitive coeducational contexts such as high-pressure standardized tests may magnify gender differences in performance. Therefore, the

class was designed to benefit all students, but to be particularly helpful to girls' who may experience stereotype threat regarding their mathematics aptitude.

During the school year, the two math workshop instructors met formally once during each eight-day cycle in the schedule to review the prior math workshop lessons and to discuss the upcoming ones. I recorded notes after each session regarding the student absences and/or interruptions to class as well as impressions about how the material was received. There were also many informal conversations between the math workshop instructors regarding how the boys and girls engaged with the material.

Evaluation Design

The evaluation of the research project used both quantitative and qualitative methods, and was based on a pretest-posttest design with multiple comparison groups. The study design was limited by the ethical requirements of fieldwork conducted in a school setting. (Rossi et al., 2004). It was not possible to conduct a randomized control trial in which some girls were randomly assigned to the mathematics workshop intervention and others were not, especially when there were reasons to believe the program might benefit the students (Rossi et al., 2004). However, it was possible to have multiple comparison groups, which are defined as untreated groups that have not been randomly assigned (Wholey, Hatry, & Newcomer, 2010). The units of a study are the level at which a treatment is assigned and evaluated (Stuart, 2007). The units in this intervention were at individual student level. Students in fifth and seventh grades received the mathematics workshop intervention and students in sixth and eighth grades did not. Therefore, it was possible to observe outcomes under both intervention and control conditions (Stuart, 2007). The evaluation design had a quantitative priority in which the collection of standardized test data received greater emphasis and qualitative data from field notes and

interviews played a secondary role (Creswell & Clark, 2011). It followed an explanatory sequential design in which quantitative data were gathered first and qualitative data were used to gain greater insight into the results of the quantitative data analysis (Creswell & Clark, 2011).

Qualitative data in the form of student interviews were gathered in a second phase to help understand the patterns seen in the quantitative data. A mixed-methods approach was suited to this study because results from the CTP-4 testing benefitted from further exploration to understand the causes for the gender discrepancies in CTP-4 mathematics scores. This second phase of data collection also strengthened reliability and validity by providing additional data sources and allowing for triangulation (Creswell & Clark, 2011).

Data Collection and Analysis

CTP-4 Standardized Mathematics Test Results

This study compared the changes in mathematics achievement scores of students in fifth and seventh grades who received the mathematics workshop with students in sixth and eighth grades who did not receive the intervention treatment in order to try to approximate the role of an independent control group. Using CTP-4 data from prior years, the growth in CTP-4 Mathematics scaled score was calculated for each student. This interrupted time-series design is considered relatively strong (Wholey et al., 2010). Within the treatment groups, the change in test scores of boys' and girls' were compared. The purpose of this comparison was to determine if the intervention benefitted girls' more than boys.

Modified Mathematics Attitude Survey

For the intervention year, the mathematics attitude survey used in the needs assessment was shortened and refined in order to more specifically target the constructs of self-efficacy and sense of belonging. It was also shortened to increase the chance that students would complete it

thoughtfully. The Pre-Survey was given to students in September 2016. All students took the survey on an individual school iPad during their regular mathematics class. The link to the survey was posted on Google Classroom. Next, each middle school teacher said the following out-loud:

> Thank you for taking part in this survey. We are trying to find out more about how best to teach math in middle school. Today we have some questions we would like you to answer about your experience in math classes. There are no right or wrong answers to any of these statements; we are interested in your honest reactions and opinions. Please read each statement carefully and indicate the number that reflects how much you agree. Your responses will be kept confidential.

The survey data collected using SurveyMonkey and exported to SPSS for analysis. All questions were randomized except for demographic questions in order to limit the potential priming of stereotype threat. Answers to the survey questions were coded as follows: “strongly disagree”=1, “disagree”=2, “agree”=3 and “strongly agree”=4. Gender was coded as 1=female and 2 =male. Grade level was coded as “fourth grade”=1 ,”fifth grade”=2, “sixth grade” =3, “seventh grade”=4 and “eighth grade”= 5. Sub-scale index scores for self-efficacy and sense of belonging were calculated for each respondent. The survey was analyzed for internal reliability. Cronbach’s alpha for the revised survey containing 12 items α .873 indicating a high level of consistence among the items. In May 2017, the same survey procedure was repeated.

The differences between the pre-intervention and post-intervention responses were calculated for each student. The survey data were then merged with standardized testing data to test for relationships between students’ self-efficacy and achievement as well as sense of belonging and achievement. T-tests of independent means were used to compare means for male

and female students on questions regarding self-efficacy and sense of belonging. Responses from students who were enrolled in an advanced mathematics class were also compared to those who were enrolled in a grade-level class, to see if the intervention differentially affected the highest achieving students.

Classroom Observations

Classroom observation data were gathered using the Reformed Teaching Observation Protocol (RTOP). After the observations were conducted, the results were scanned and saved on a computer. The data were transferred into Microsoft Excel so that basic statistical analyses could be conducted. A mean score on each section of RTOP was calculated for each teacher for each subtopic (lesson plan, content, and classroom culture), as well as a mean score for the entire observation instrument. These scores were compared to each other in order to assess the degree to which the two math workshop teachers used similar pedagogical techniques in the classroom.

Open-Ended Interviews

The results from the twelve interviews conducted by Director of Research were recorded using the application Dictate and Connect. The audio files were emailed to me, and I saved them in a secure Dropbox folder. Next, the audio files were uploaded to a private folder where they were accessed by a professional academic transcriber. She returned complete transcriptions and then deleted all of the audio files from her computer.

The interview analysis process was informed by the work of Green et al. (2007) on how to generate the best evidence in qualitative data analysis. The first step in analyzing the data was "immersion" in the data through repeated reading of the transcripts. This allowed for making connections among interviews and the listing of possible themes. Next, the transcripts were coded using the qualitative software NVivo for Mac for the major themes from the theoretical

framework: gender saliency, sense of belonging and self-efficacy. The third step in the interview analysis was linking of codes into categories and relationships. The last step in the analysis was the generation of themes. A theme in the data was not simply a category; instead, it was an interpretation or an explanation for the patterns in the data that was linked to theory.

Strengths and Limitations of Study Design

The combination of a quasi-experimental design with multiple comparison groups and an interrupted time series design is the most powerful of all quasi-experimental designs (Shadish, Cook, & Campbell, 2000). The use of a pretest provided a limited ability to infer what would have happened to the girls' scores if they had not participated in the intervention (Shadish et al., 2000). However, these inferences are weak at best because of multiple threats to validity including selection, maturation, and history (Shadish et al., 2000). In the case of the math workshop intervention, students were simultaneously enrolled in their regular, daily mathematics classes. Thus, it is likely that any growth in mathematics test scores could be attributed to learning that took place in that setting or any other number of educational experiences. Students are projected to make annual progress on the CTP-4 mathematics test; therefore, a treatment effect must be considered only after accounting for expected maturation first.

Including multiple comparison groups strengthened the evaluation design. A comparison group design can be used to assess program impact if certain conditions are met (Wholey et al., 2010). One concern with comparison group design is that two groups that are compared will suffer from selection bias. For this evaluation design, selection bias between grades was minimized by the fact that all fifth and seventh grade girls were required to participate in the treatment. Furthermore, the grades that participated in the mathematics workshop were chosen due to scheduling convenience and not due to any difference on any of the research variables.

Thus, there was limited concern that the treated classes were selected for some reason that made them different from the control group.

There were additional threats to validity that must be considered when comparing the male and female classes within the same grade. The scores for boys and girls on the pretests differed considerably, thus there was the chance that selection combined with other threats to validity (Shadish et al., 2000). In order to rule out selection bias, one must be sure that if the treatment had not been offered, the annual gain in achievement on standardized tests measures would be expected to be roughly the same for females and males (Wholey et al., 2010). This cannot be assumed, as projected gains differ depending on the place in the achievement distribution and on average boys began at a higher scale score than girls. On a normed achievement scale, getting a single item correct has more effect on performance for those at the extremes of the distribution than at the mean (Shadish et al., 2000). There is also the possibility of selection-history, the fact that an event occurred between the pretest and posttest that affected either the boys or girls differentially (Shadish et al., 2000).

This study design was also threatened by a lack of stability (Stuart, 2007). In the school setting, it was not possible to ensure that individual students' outcomes were not influenced by other students. Students routinely interacted inside and outside the classroom, and the social dynamics created could have influenced the outcomes of the study (Stuart, 2007). In addition, by necessity there was more than one version of the treatment—boys' sections had a different instructor from the girls' sections. The only means to control for this was to accept a set of activities as variations of the treatment program (Stuart, 2007). Finally, there were a number of threats to construct validity that must be addressed. These included reactive self-report changes, reactivity to the experimental situation, novelty and disruption effects, and experimenter

expectancies. For example, seventh grade students in the pilot study had highly emotional reactions to the creation of single-sex mathematics classes. An emotional response may influence students' self-report responses on surveys and during interviews (Shadish et al., 2000).

Projected Effect Size

Studies on the effect of all-girl mathematics classes on mathematics achievement tests have shown only small to moderate effect sizes by conventional benchmarks (Cherney & Campbell, 2011; Eisenkopf et al., 2015). However, Lipsey et al. (2012) propose that effect sizes benchmarks must be re-calibrated for educational research. Specifically, Lipsey et al. (2012) argue that an effect size on a mathematics achievement tests as large as .30 is rare. Thus, for educational research using standardized tests of performance, the authors suggest that an effect of .25 be considered large. Furthermore, the median effect size for a standardized test with a narrow scope for middle school students is .26 and the mean effect size is .32 (Lipsey et al., 2012). In one study of the effects of the Teach for America program on students' academic achievement in mathematics and reading, Decker, Mayer, and Glazerman (2004) found an effect size of 0.15 on mathematics achievement test scores. Studies on the impact of all-female mathematics classes have found results considered moderate to large using this benchmark.

One randomized study on achievement differences for females in single-sex and coeducational mathematics classes found an effect size of (d=.17-.24; Eisenkopf et al., 2015). The potential sample population of the intervention may not be large enough to detect an effect of this size. In order to reliably detect an effect size of .25 with 0.8 power, it would be necessary to have a sample size of at least 128 participants. The pilot year had 123 students receiving the pilot math workshop and approximately the same number who did not. Furthermore, only

students whose parents agreed to have their data included in this research study were part of data analysis.

Conclusion

This chapter discussed the mathematics workshop intervention procedure and the methodology for evaluating this intervention. It began by reviewing the problem in context and the primary research questions. It then provided information about the research participants, the research setting, and the researcher's role as a participant-researcher. It reviewed the qualitative and quantitative data sources that were gathered during the pilot year, the pilot year methodology, and how the pilot year experiences were used to change and adapt the program year design. Finally, it discussed the intervention methodology and the strengths and limitations of the research design as well as the effect size that could be expected in this type of intervention. Chapter 5 will discuss the implementation and results of the pilot year; Chapter 6 will discuss the results of the intervention year.

CHAPTER 5: PROGRAM IMPLEMENTATION

Chapter 4 discussed the intervention procedure, data collection, and analysis methodology. This chapter will report on the pilot year implementation of the math workshop. The work of Rogers (2003) is used to understand the process by which an innovation is adopted or fails to be adopted by a community. The process of the "diffusion" of a new idea involves communication and social change. This chapter will discuss the diffusion process in the middle school where this pilot program was implemented. It will review how the school community first responded to the intervention; potential explanations for these responses, and changes that were made to the program to better align it with the school's values and needs. This chapter will also discuss the pilot year results, and how the totality of the pilot year experiences influenced the design of the second year of the program.

Pilot Year Implementation

The community reaction to the creation of single-sex math workshop was complex, and Rogers' (2003) work on the diffusion of innovations can be helpful in making sense of the dynamics associated with the adoption of this intervention. Using Rogers' (2003) definition of an innovation as "an idea, practice, or project that is perceived as new by an individual or other unit of adoption" (Rogers, 2003, p. 12), the mathematics workshop program was an innovation because it was perceived as a new program to the members of the school community. Importantly, Rogers (2003) proposes that some perceived attributes of an innovation predict a quick rate of adoption while others do not. A first key attribute of an innovation is its perceived relative advantage, or the degree to which potential adopters believe that an intervention is significantly better than what it is replacing. The majority of students in the pilot year did not readily agree that an advantage would be conferred in having an all-female class. Indeed, the

research supporting the merits of single-sex classes is nuanced and even suggests that there could be disadvantages to all-female classes. Thus, it makes sense that students in seventh grade would have a hard time making sense of the change. Adults in the community had an easier time anticipating the opportunities for advantages of the intervention, although some did not find the research convincing or were pedagogically committed to coeducational classrooms.

Compatibility with an institution's values and needs is a second attribute that predicts an innovation's rate of adoption (Rogers, 2003). Given that the BC School is a coeducational school, on the surface the intervention conflicted with the school's foundational identity and commitment to educating boys and girls together. However, there were other elements of the school's culture and mission that were highly compatible with the intervention. First, the school is highly devoted to social justice, and the faculty is motivated to address inequalities in educational outcomes whenever possible. Additionally, the school leadership prides itself on being innovative and progressive in applying the latest research-based advances in the education field. These two elements of the school's identity were consistent with adoption of an intervention of a single-sex math workshop in order to address the school's urgent need to address the gender gap in mathematics scores and attitudes. The introduction of the math workshop was also not entirely new this year; in 2014 to 2015, there was a supplementary mathematics class offered for sixth grade students. As Rogers (2003) notes, the rate of adoption may be sped up if it is part of a group of innovations that are introduced sequentially.

A third perceived attribute of an innovation is complexity. In general, the more difficult or complicated an innovation is to use, the slower the rate of adoption (Rogers, 2003). The logistics of implementing a single-sex mathematics section were relatively simple and could be considered low in complexity. The intervention did not require much training, new materials, or

resources—students were simply split into two groups based on gender identification. However, the *ideas* behind the implementation were extremely complex. Faculty, parents, and especially some students struggled with understanding why and how an all-girls mathematics class could help improve girls' confidence and achievement in mathematics. Interviews with students who strongly disagreed with the gender division indicated that they held misunderstandings about the nature of stereotype threat. Some felt that they had never heard of the negative stereotype about girls in mathematics and/or that the discussion about this stereotype strengthened instead of weakened the power of it to disrupt girls' performance. Effectively communicating and leveraging the evidence-based rationale for single-sex mathematics classes was the most persistent challenge in gaining support for the intervention.

Finally, although the intervention of a single-sex math workshop was relatively straightforward, the ability for others to observe results attributable to the innovation was more difficult. Rogers (2003) suggests that perceived observability of an innovation is positively correlated with its speed of adoption, and that an innovation entirely composed of ideas has a lower degree of observability. Students' attitudes may not substantially change in only twenty sessions of math workshop, and if these attitudes do change, they may not be easily observed in daily behavior. To improve the observability of the results for the researcher and community alike, data regarding attitudes and achievement in mathematics were collected, analyzed, and shared.

The needs assessment informing the pilot year was conducted in April 2015. Following analysis, interpretation, and write-up of those results, the math workshop program was introduced to the parents and teachers in the community in August 2015 at the beginning of school during the pilot year. This pacing was tight and may have contributed to some of the

adoption challenges. Because the "adoption" of the math workshop program was mandated only a few weeks before the beginning of school, students, faculty, and parents did not have the opportunity to complete all the steps in the innovation- decision process outlined by Rogers (2003). A letter to middle school parents informing them of the math workshops was sent only a week prior to the start of school. It explicitly mentioned the school's concern over a gender gap in mathematics test scores, which mirrored gender gaps seen across the country. Thus, their knowledge of the innovation occurred at the same time as the implementation. There was little time for individuals to form opinions and communicate about their attitudes towards the innovation. Faculty outside of the mathematics department received information regarding the math workshops at virtually the same time as the parent. A middle school faculty meeting was scheduled in late September 2015 to establish the need for the intervention and to provide information about the math workshop. Once the middle school faculty became aware of the lower mathematics achievement and negative attitude of girls towards mathematics, they expressed a greater understanding for a need to change. Time is an important component of the innovation-decision process. In this case, the limited time allotted to individuals for information gathering and processing may have been a challenge to widespread adoption.

Students in fifth and seventh grades who participated in math workshop were not officially informed of the change until they arrived at the first class. Students may have learned from their parents about the initiative before the start of school when their parents received the email about the program. However, this information could also have been filtered through the positive or negative perceptions of the parents. During the first two sessions of math workshop, the reasons behind its creation and implementation were discussed with the students. The Director of Inclusion (both female) and I facilitated the girls' sections in both fifth and seventh

grade while the male Math Department chair and the Director of Innovation facilitated the boys' grade sections because the girls' and boys' sections met concurrently.

As Rogers (2003) discusses, an intervention can be framed in various ways and these choices are consequential for how the intervention is perceived. In this case, the goal was to frame the math workshop as a means for addressing the problem that girls are subject to the negative stereotype that boys are better at mathematics. The facilitators discussed the definition of a stereotype and the fact that currently girls and women are underrepresented in mathematics and sciences. They suggested that practicing mathematics in an all-girls environment could sometimes help girls to build confidence and enjoyment in mathematics. Despite attempts at clear and age-appropriate framing of the innovation, many students believed the class was a remedial mathematics course for girls. The content of the email sent to parents highlighting the gap in test scores may have contributed to the students' insistence that the course was intended as a test-taking remediation for female students. As Rogers (2003) emphasizes, diffusion is a social process, and individuals often perceive innovations in ways that are not intended, desired, and/or expected by the change agents.

One challenge given the mandate of adoption was identifying the level of support for the workshop among faculty, parents, and students. Essentially, the information-decision process was completed concurrently with the implementation of the innovation. As Rogers (2003) observes, during the innovation-decision process individuals reduce their uncertainty about the innovation by seeking information. Many members of the community did not have the opportunity to reduce their uncertainty about the innovation to a level that would allow them to unequivocally adopt the idea. Unfortunately, the period of time in which change to an innovation can occur is usually limited (Rogers, 2003). The pilot year was a critical phase of this

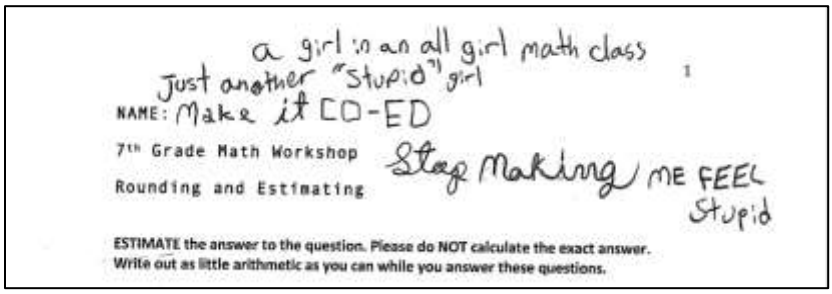
intervention in which misalignments between the school and the innovation needed to be overcome in order for the program to be improved for the following year. In that sense, one lesson from the pilot year may be that it should have been framed, as the beginning of an extended decision making process that would last a minimum of two years. This timeline would have been consistent with study and dissertation requirements while also allowing the community to engage more naturally in a developing discussion.

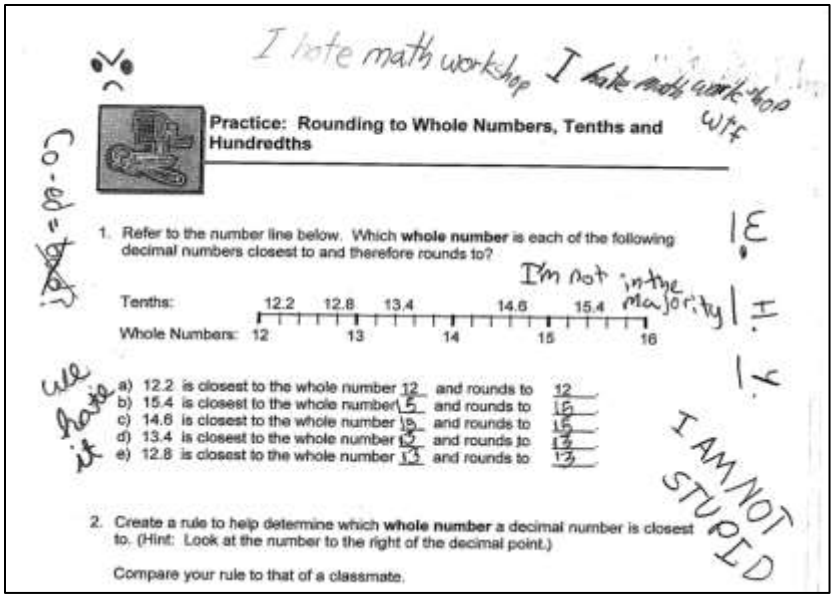
Student Response to the Pilot Year

Students in fifth grade seemed largely indifferent to the division by gender. After the first class, the fifth grade students rarely mentioned this feature of the course or made any explicit positive or negative comments regarding the gender division. However, during the pilot year, the seventh grade girls had a very different reaction. From the initial meeting, girls in seventh grade expressed a high degree of outrage about the single-sex arrangement. As Figure 10 depicts, several girls wrote angry comments on their papers expressing their frustration. These comments help elucidate why students felt angry about math workshop. First, the comment, "I am not stupid," and "stop making me feel stupid," suggests that the student interpreted the school's choice to create single-sex classes as a confirmation that school faculty and administration believed girls are not as intelligent as boys, in short, confirming the negative stereotype instead of combating it. By speaking of this stereotype openly, the students felt the school had made it more powerful and damaging. Psychological research on the value of discussing stereotypes openly is mixed. While some research proposes that teaching about stereotype threat may lessen its power, (Johns, Schmader, & Martens, 2005), there is also existing literature that suggests teaching about stereotype threat could exacerbate the problem it describes by priming negative

thoughts about the stereotype that is discussed (Johns et al., 2005). The mixed nature of the student reaction seems to mirror the inconclusive nature of the published research.

Another written comment, "I am not in the majority," indicates that this girl may have felt that the reasons given for the single-sex section were not applicable to her personally, a feeling that was also reflected in the written comment "this is a real stereotype but it isn't in our grade." These comments suggest that the students rejected the notion that they may be subject to stereotype threat and find the implication that they may be influenced by negative stereotypes about women in mathematics to be offensive. Interestingly, there is research suggesting that as a coping mechanism, members of a disadvantaged group may adopt extreme meritocratic beliefs and refuse to view themselves as targets of prejudice (Barreto & Ellemers, 2005). This belief that the existing social structure is legitimate and based on one's actions and abilities may allow individuals to maintain that the world is predictable and under one's control (Schmader, Johns & Barquissau, 2004). In their study of "old-fashion" more overt vs. "modern" sexism, Barreto and Ellemers (2005) found that female participants were less likely compared to male participants to recognize more subtle forms of sexism such as denial of gender discrimination and resentment of women's demands. Essentially, the seventh grade girls may not have wished to acknowledge that they are members of a disadvantaged group, and this may have led to their becoming less able to recognize and challenge prejudice expressed in subtle ways.





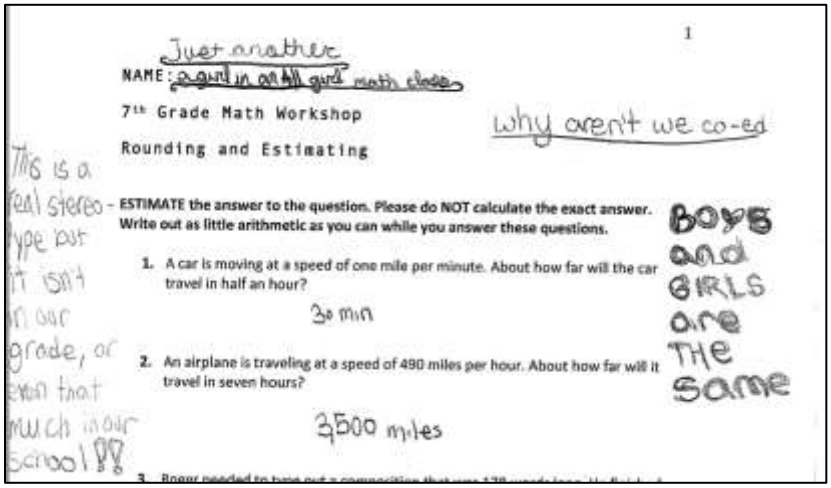


Figure 10. Examples of comments on classwork of seventh grade girls in math workshop, 2015-2016.

To express their discontent further, the seventh grade students wrote a lengthy signed petition to the school faculty. The petition had several key points that echoed the concerns of the angry comments initially written on papers. First, despite a detailed explanation for why the program was implemented, the seventh grade students felt the program implied girls were not as capable as boys in mathematics—and essentially confirmed instead of denied this stereotype. The petition read, "This class makes girls and boys alike feel as though you believe girls are not as good at math as boys." Despite many efforts to dissuade students that the school administrators and faculty believe females are equally capable in mathematics—and this belief was precisely the reason this class was created-- many students remained unconvinced.

Second, the students did not believe in the idea of stereotype threat as faculty had presented it. They claimed they were unaware of their gender in mathematics class and thus were not susceptible to stereotype threat. In the students' words, "When we are in a classroom situation that is coed, we don't spend time pondering our gender. In fact, we don't think about our gender at all." This quotation revealed a misunderstanding about the very nature of stereotype threat. The students had trouble understanding that stereotype threat refers to subconscious and implicit beliefs and not necessarily to a student's conscious awareness of identity threats. Interestingly, girls with the highest mathematics achievement scores were the most upset about the gender division. Additionally, because mathematics classes in the school were "tracked" girls in more advanced mathematics classes were not accustomed to be in class with girls in less advanced classes. The petition read, "If a girl struggles in math, she can request help, but to bring the rest of the girls down is unacceptable." This quotation indicates a deep resentment among the girls who signed the letter for having to participate in a class with peers who may have found mathematics more challenging. Research on stereotype threat suggests that

girls who are most talented and interested in mathematics are most vulnerable to its forces (Inzlicht & Ben-Zeev, 2000). It may be noteworthy that it was these high-ability girls who were most angry about the single-sex groupings in the pilot year. Their anger suggested a high degree of investment in their identity as strong students and mathematicians.

Finally, seventh grade student also complained that math workshop "forced students into set binary boxes of boy or girl," and that this wrongly "encourages the idea that there are only two options of gender, male and female." This posed a real challenge to the intervention, as the school openly supports and encourages students to self-identify their preferred gender identity and personal pronouns. The desire to provide an affinity space for female students appeared to be in conflict with the motivation to ensure all students felt they were able to be anywhere on the spectrum of gender identity. All of these concerns regarding the math workshop intervention were important to consider when deciding how to change and adopt the program for the following year. Rogers (2003) argues that the re-invention and adjustment phase of a new innovation is crucial in determining its success or failure to be adopted.

Understanding the Pilot Year Response

Understanding the specific school context also provides information to help understand the strong reaction to the intervention. This BC School places a high value on curricula that addresses social justice and the encouragement of student as social activists. In particular, there is a social justice curriculum in sixth grade that focuses on questions of human dignity. Many seventh grade students drew on their experience in that class to claim that the math workshop intervention did not respect human dignity because it called on an exterior factor (gender) to divide individuals into groups. In fact, the same students extended this idea to argue the school's students of color affinity group was also an example of "tribalism" and should not continue. The

term tribalism generally refers to "behaviors or attitudes that stem from strong loyalty to one's own tribe or social group" (English Oxford Living Dictionaries, 2017). In the sixth grade curriculum shared vocabulary document, tribalism was presented as "distrust, fear, and vilification of people who are not normally included in the "Circle of Us" (Common Vocabulary-Dignity). During the pilot year, seventh grade students perceived the division of boys and girls in math workshop as negative tribalism and felt that it directly contradicted their social justice curriculum. After discussion with the sixth grade teachers, the following year the curriculum was adjusted to include a discussion of when tribalism could also be positive in motivating "feeling of community that links individuals together" (Common Vocabulary-Dignity). Students in fifth grade had not yet participated in this social justice class, and may not have been primed to see these issues through the same lens. Based on the student petition, it appears that students made sense of some of the complex concepts in a manner that led to rejection of any formation of groups based on shared experience.

A more developmental hypothesis for why students in fifth and seventh grade would have such different responses to the program is the gender intensification hypothesis (Frenzel, Goetz, Pekrun, & Watt, 2010). In fifth grade, many students are beginning puberty and the associated identity-building process. By seventh grade, gender socialization and gender intensification may peak (Frenzel et al., 2010) resulting in a higher response to interventions that target gender salience. Thus, although seventh grade students had a stronger negative reaction to the intervention, it may nevertheless be that this age group will benefit more from an intervention that targets gender salience.

Pilot Year Results

The interpretation of achievement test data in 2016 was complicated by a change from paper-based testing in the middle school to iPad-based testing. Furthermore, the new digital version of the test is a computer adaptive test by section. A students' score on the first half of the achievement test determines if they received either an easier or more challenging second half of the test (CTP Technical Report, 20154). However, the ERB, which administers the CTP-4, claims:

> The second-stage sections within a content area and level are not distinctly different tests. In many cases, a degree of item overlap exists between the second-stage sections. In other instances, items are not repeated between the second-stage sections but a number of items share very similar statistics with regard to difficulty and discrimination. Consequently, students are not taking different tests per se, but rather they are being administered tests that more appropriately enable them to perform to the best of their ability. (ERB, p. 19)

In order to allow comparison between the two versions of a test ERB created a new vertical scale of scores was created so that paper-based tests and online tests were linked to the same scale. Despite this fact, it is important to note that changing the modality of test delivery might have some unknown effects on student test scores and that comparisons between the paper-based scores and the online test scores during this one year of transition may be less reliable.

In April 2016, middle school students took the CTP-4 battery of tests. When interpreting student achievement growth, it is helpful to know that ERB advises schools that students should gain on average about 7-10 scaled score points per year. However, ERB also acknowledges that this may vary depending on the prior achievement of the student because it is more difficult to move up the vertical scale at the very top of performance (ERB, 2014). In 2016, scores growth

was larger for fifth grade students because this cohort was last tested in fall 2014 with a grade 3 test instead of spring 2015 with a fourth grade test. On the Mathematics subtests, fifth grade boys gained approximately two more scale score points on average than fifth grade girls, but this difference was not statistically different by gender (Table 15). The same pattern was seen when considering only students in advanced mathematics sections, with boys gaining slightly more than girls (Table 16).

Seventh grade girls' scores increased on the Mathematics achievement test more than boys did, and the gender gap was reduced by approximately 6% (Table 17). However, both groups gained relatively few scaled score points. When the same analysis of means was run selecting only for students in advanced mathematics classes, the girls' gain was more pronounced but still below statistical significance (Table 18). Seventh grade boys gained on average less than one scaled score point. This may be partially explained by their being a strong cohort who were already high-scoring and therefore at a disadvantage for improving their scaled scores due to ceiling effects. Although the difference in test score gains did not reach statistical significance, it seemed noteworthy in the context of the strong resentment of the program among seventh grade girls and the rushed implementation. The school administration felt that even this modest reduction in the gender gap was encouraging enough for the administration to renew the program for a second year with some adjustments.

Table 15

2016 Results of t-test for Change in Mathematics Achievement by Gender in Fifth Grade Math Workshop Intervention

	Gender								
	Boys			Girls			95% CI for Mean		
	M	SD	n	M	SD	n	Difference	t	df

Change in Scaled score - mathematics	23.06	17.21	31	21.44	14.06	25	-10.18, 6.94	-3.80	54

p=.705

Table 16

2016 Results of t-test for Change in Mathematics Achievement by Gender in Fifth Grade Math Workshop Intervention for Students in Advanced Math Section

	Gender						95% CI for Mean Difference	t	df
	Boys			Girls					
	M	SD	n	M	SD	n			
Change in Scaled score - mathematics	21.94	17.86	18	17.44	16.70	9	-19.21, 10.21	-.630	25

p=.705

Table 17

2016 Results of t-test for Change in Mathematics Achievement by Gender in Seventh Grade Math Workshop Intervention

	Gender						95% CI for Mean Difference	t	df
	Boys			Girls					
	M	SD	n	M	SD	n			
Change in Scaled score - mathematics	1.56	10.29	34	4.94	12.21	35	-2.05, 8.82	1.24	67

p=.218

Table 18

2016 Results of t-test for change in Mathematics Achievement by Gender in Seventh Grade Math Workshop Intervention for students in Advanced Sections

	Gender						95% CI for Mean Difference	t	df
	Boys			Girls					
	M	SD	n	M	SD	n			
Change in Scaled score - mathematics	0.687	9.61	32	5.77	3.391	22	-1.19, 11.36	1.62	52

p=.110

Pilot Year Survey Results

Analysis of survey results from spring 2016 was limited by a low participation rate. Only 60 out of the 123 students (49%) who participated in the math workshop pilot program completed both the 2015 and 2016 mathematics attitude surveys. In fifth grade, there were 14 girls who completed both survey and 10 boys. In seventh grade, there were 20 girls who completed the surveys and 16 boys. Furthermore, the survey was only given to students in fifth and seventh grades who had participated in the math workshop program that year, which limited the ability to compare the results with other students who had not received the pilot program. In the fifth grade, low completion rates were partially explained by the fact that students in fifth grade for 2015-2016 took the survey as part of the needs assessment as lower school students during their fourth grade year in a separate building and without the same familiarity with iPads. In seventh grade, low completion may be due in part to the anger that some students felt regarding the program.

None of the changes in self-efficacy scores or sense of belonging scores were statistically significant by gender. However, there were some general patterns in the data worth noting. First, as expected, boys in both grades self-report higher levels of both self-efficacy and sense of belonging. Second, the scores for students' self-reported self-efficacy between 2015 and 2016 did not change meaningfully. In contrast, there was an increase in both boys' and girls' reported sense of belonging for fifth grade students.

Table 19

Mean Self-Efficacy and Belonging Spring 2015 (Needs Assessment)

Grade	Mean Self-Efficacy		Mean Sense of Belonging	
	Girls	Boys	Girls	Boys

Rising fifth	2.68	3.13	2.71	2.90
Rising seventh	2.76	3.13	3.30	3.34

Table 20

Mean Self-Efficacy and Belonging Spring 2016 (Pilot Year)

Grade	Mean Self-Efficacy		Mean Sense of Belonging	
	Girls	Boys	Girls	Boys
Fifth	2.69	2.97	3.25	3.30
Seventh	2.77	3.18	3.15	3.43

Reinvention of the Math Workshop Program

A number of important changes were made to "reinvent" math workshop for the second year and increase the chance of its adoption, which Rogers (2003) defines as "the full use of an innovation as the best course of action available" (Rogers, 2003, p.177). Reinvention is the degree to which an innovation is changed or modified by the user during the process of implementation, and innovations that are reinvented are more likely to be sustained over time (Rogers, 2003).

As discussed in Chapter 4, reinvention of the math workshop included changes to both the curriculum and pedagogy. The researcher and several faculty members from the mathematics department met over the summer for additional time to plan revised math workshop lessons. First, the curriculum and pedagogy of the math workshop program was revised in order to a shift away from individual and teacher-focused instruction towards open-ended problems that students would complete collaboratively. Problems were selected that were "low floor / high ceiling"

meaning that they could be approached at many different levels of a mathematics from a more basic approach of "guess and check" to formal algebraic models (see example lesson Figure 2). The pilot year indicated that some students might respond negatively to the intervention, and so it was important to engage all students as quickly and completely as possible in a task that they viewed as interesting and enjoyable.

The Situation: The FBI recovered this enormous stack of money stolen by international thieves!

Your Challenge:

How much money is that?!

Questions To Ask:

- What is a guess that is too low?
- What is a guess that is too high?
- What options do you have for counting this money and what are the advantages and disadvantages to each method?
- How can we measure the volume of $100 bills in the pile? How important is it that we have an accurate answer?
- What is your best guess?

Figure 11. Example of math workshop activity.

A second important change was a shift in communicating about the reasoning and research behind the program primarily with *parents and faculty* instead of students themselves. As Rogers (2011) notes many times, adoption of an innovation is primarily a social process that requires careful communication. In 2016 to 2017, fifth and seventh grade students participating in math workshop were introduced to the course with a brief one-sentence explanation that "sometimes doing math in a single-sex group helps girls feel more confidant and have more fun doing math." Following this introduction, students were immediately engaged in a collaborative problem-solving task. In order to better engage adult stakeholders in the community, faculty participated in this same problem-solving task in teams during a faculty meeting in September, and parents were invited to an evening in which they too took part in an example activity. The goal was to increase community adoption of the program by communicating directly with faculty and parents and demonstrating the types of activities that students would be doing. Both sessions appeared successful as adults reported enjoying the activity and had an opportunity to ask questions and meet the math workshop teachers.

A third change was that students in the second year were asked to self-identify their gender identity at the beginning of the year. If their gender identity did not match their assigned sex at birth they would be asked which class they felt most comfortable attending. In 2016 to 2017, there were no students who identified a gender that did not match their gender assigned at birth. However, the goal was to provide students a greater sense of choice in the process of being assigned to either a girls' or boys' section.

The combined changes that were made to reinvent the math workshop were successful in increasing the acceptance and adoption of the program by the school community. Students

during the second year did not express a similar level of anger about the creation of single-sex groups. On the end-of-year reflections that students filled out, the few complaints focused on having an additional mathematics class in the schedule or feeling that it was not "real math," because the activities were inquiry-based. However, none of the students reiterated any of the main arguments that seventh grade students documented in the pilot year petition.

Conclusion

This chapter discussed how the school community responded to the math workshop program during the pilot year and changes that were made to reinvent the intervention to further its success and adoption for the following year. This narrative used the work of Rogers (2011) on the diffusion of innovations as a means for making sense of what perceived attributes of the math workshop initiative may have helped and hindered its successful adoption. This chapter also reported on the results from the pilot year, and changes that occurred to the instruments of data collection including the mathematics attitude survey and the CTP-4 standardized testing administration. Chapter 6 will report on the results of the intervention year and discuss the implications of these findings.

CHAPTER 6: RESULTS AND DISCUSSION

The purpose of this dissertation was to understand how high-ability middle school students' attitudes towards mathematics and achievements in mathematics differ by gender and to report on an intervention to address the gender gap in mathematics achievement among high-ability students at an independent middle school. In Chapter 5, the pilot program implementation and results were discussed as well as changes that were made to the program for the second year. The chapter will discuss the process evaluation and results for the second, intervention year of the math workshop program. As discussed in greater detail in Chapter 4, the intervention year consisted of 20 supplementary single-sex mathematics classes for students in fifth (n=65) and seventh grade (n=71) at a coeducational middle school. Although the results of the full data set were analyzed for the school's program evaluation, the results reported here will include only those students for whom written parental permission was obtained.

The results of the intervention were measured with three primary data sources: standardized achievement test scores, a pre and post intervention survey, and open-ended interviews with 12 randomly selected participating girls. This mixed methods research followed an explanatory sequential design. Thus, when discussing data, the quantitative data will be discussed first and qualitative data from interviews will be used to help understand and explain the quantitative findings to suggest new directions for future research.

The results will be discussed for each of the guiding ,s:

RQ1: Will middle school students' participation in supplementary single-sex mathematics classes benefit girls more than boys as measured by sense of belonging, self-efficacy, and achievement?

RQ2: Will a change in girls' self-efficacy correlate with a change in mathematics achievement test scores?

RQ3: Will a change in girls' sense of belonging correlate with a change in mathematics achievement test scores?

Process Evaluation: Fidelity of Implementation

Fidelity of implementation refers to the degree to which a program is implemented in the way that was intended by the program developers (Dusenbury, Brannigan, Flaco, & Hansen, 2003). After the math workshop program was revised and communicated, the delivery of the program was documented during the intervention year to assess the fidelity of implementation. For this research project, fidelity was measured using the following five components: level of adherence to the program, dosage, quality of delivery, participant responsiveness, and the level of program differentiation (Dusenbury et al., 2003).

All of the dimensions of fidelity of implementation were assessed with multiple measures to increase reliability. These measures include data gathered by the math workshop instructor, outside observers and student participants. The instructor kept a running log in the form of field notes and gathered artifacts from class. Three trained school administrators conducted observations using the Revised Teacher Observation Protocol (RTOP). This classroom observation tool uses a five point Likert scale (0-4) to assess the lesson design and implementation, lesson content, and classroom culture (Sawada et al., 2002). Twelve students were randomly selected and interviewed at the end of the school year and all participating students filled out an anonymous reflection during the last lesson.

The level of adherence to the planned math workshop lessons was high. The two instructors shared lesson plans and materials on Google Drive. Instructors met weekly to discuss

how the previous lesson went and to review the upcoming lesson. In every lesson, the girls and boys sections completed the same activity and used the same materials. The planned dosage of the workshop was 20 sessions over the course of the academic year. However, only 15 of those sessions occurred before the CTP-4 mathematics achievement tests in April 2017. Furthermore, due to field trips and a snow day, each section missed approximately two full sessions. Thus, the dosage of the intervention was lower than anticipated with approximately 13 full sessions before the standardized testing occurred in April.

The quality of the delivery was assessed by teacher observation. Administrators who had been trained in using the RTOP protocol observed one instructor on three different occasions and the other instructor on four occasions. For each statement included in RTOP, the observer rated the math workshop teacher from 0 (not present at all) to 4 (very descriptive). The scores from these observations were averaged to give each instructor an average score in each RTOP domain as well as an average total score. The average teacher scores on the RTOP protocol were nearly identical (80.3 and 80.5 respectively) as were the patterns of their average scores across the different domains. For example, both instructors had classroom culture as the area on which they were scored the highest. These results suggest that that both math workshop instructors maintained similar high quality of delivery of the materials that was would be described as student-centered and inquiry-based.

Table 21

Comparisons of Math Workshop Instructor Mean RTOP Scores

RTOP Domain	Lesson Planning	Content	Classroom Culture	Total
Instructor A (Girls)	3.0	3.1	3.4	80.3
Instructor B (Boys)	3.2	3.1	3.4	80.5

Participant responsiveness in the program was assessed using teacher observation, anonymous program reflections at the end of the year and student interviews. Overall, during the pilot year, students reported enjoying the math workshop. Nine out of the twelve students interviewed called the program "fun" and all twelve recommended keeping the program in place for the following year. The process of program differentiation and understanding what components of this program are effective is ongoing. Although the single-gender nature of the intervention is the most obvious differentiator, how gender composition interacts with the curriculum and content of the program is not yet fully understood. The results of this research study suggest that it is important to provide a learning opportunity that feels collaborative and accessible to all students and to generate an atmosphere of enjoyment around playing with mathematics.

Results

Achievement Testing Results

Did middle school students' participation in single-gender mathematics classes benefit girls more than boys as measured by achievement? Analysis of CTP-4 mathematics tests results suggest that the math workshop program was effective in reducing the gender gap in mathematics test scores for some students. A 3-way between groups ANOVA was conducted to

compare the main effect of gender (boy, girl), track (on-level, advanced) and intervention (treatment, control) on the change in CTP-4 Mathematics achievement scores. There was a significant 3-way interaction, $F(1)=7.428$, $p=.007$, $\eta 2 = .051$.

To better understand these results, the change in mathematics scaled scores of sub-groups of students were compared using independent t-tests. There was no statistically significant difference in change in mathematics achievement scores in fifth grade by gender (Table 22). When students in advanced mathematics classes were selected, girls on average gained more scaled score points than boys, but not at a statistically significant level (Table 23).

Table 22

2017 Results of t-test for change in Mathematics Achievement by Gender in Math Workshop Intervention for Fifth Grade Students

	Gender						95% CI for Mean Difference	t	df
	Boys			Girls					
	M	SD	n	M	SD	n			
Change in Scaled score - mathematics	12.61	17.90	18	13.54	9.22	24	-9.51, 7.64	-2.19	24

P=.828

Table 23

2017 Results of t-test for change in Mathematics Achievement by Gender in Math Workshop Intervention for Fifth Grade Students in Advanced Math Sections

	Gender						95% CI for Mean Difference	t	df
	Boys			Girls					
	M	SD	n	M	SD	n			
Change in Scaled score - mathematics	7.88	10.93	9	15.17	9.73	12	-16.74, 2.18	-1.61	19

P=.115

In seventh grade, on average girls gained more scaled score points than boys, and the effect approached statistical significance (Table 24). When seventh grade students in advanced classes

were compared, the effect was more pronounced but again below statistical significance (table 6.4).

Table 24

2017 Results of t-test for change in Mathematics Achievement by Gender in Math Workshop Intervention Seventh Grade Students

	Gender								
	Boys			Girls			95% CI for Mean		
	M	SD	n	M	SD	n	Difference	t	df
Change in Scaled score - mathematics	6.73	13.26	26	11.07	9.73	30	-10.51, 1.84	-1.41	54

p=.165

Table 25

2017 Results of t-test for Change in Mathematics Achievement for Seventh Grade Students in Advanced Math Classes by Gender

	Gender								
	Boys			Girls			95% CI for Mean		
	M	SD	n	M	SD	n	Difference	t	df
Change in Scaled score - mathematics	3.72	13.99	18	10.25	9.50	20	-14.31, .137	-1.69	46

p=.098

There was a statistically significant result for gender when comparing the growth in Mathematics achievement scores for high ability girls and boys who received the intervention (Table 26).

Table 26

2017 Results of t-test for Change in Mathematics Achievement Test by Gender for Fifth and Seventh Grade Students in Math Workshop Intervention and Enrolled in an Advanced Math Class

	Gender								
	Boys			Girls			95% CI for Mean		
	M	SD	n	M	SD	n	Difference	t	df
Change in Scaled score - mathematics	5.11	12.99	27	12.09	9.74	32	-6.98, 2.96	-2.36*	57

* p=.022, d=.60

These achievement gains mirrored the pilot year results in two key respects—girls in seventh grade trended towards greater benefit from the intervention than girls in fifth grade, and girls enrolled in advanced mathematics classes trended towards more benefit more than girls enrolled in on-level mathematics classes. This finding resonates with the research that students in seventh grade may be more aware of their gender and experience higher gender salience than fifth grade students (Frenzel et al., 2010), and that students of high-ability and high identification with mathematics are more vulnerable to stereotype threat (Inzlicht & Ben-Zeev, 2000). Therefore, girls in late-middle school of high prior mathematics achievement and mathematics self-identity may benefit more from an all-girls mathematics intervention that reduces the deleterious effects of stereotype threat.

Survey Results

Research Question 1

Did middle school students' participation in single-sex mathematics classes benefit girls more than boys as measured by sense of belonging and self-efficacy?

Pre-Survey September 2016

In September 2016, 251 out of a total enrollment of 285 students in the middle school (89% participation rate) completed the mathematics attitude survey regarding self-efficacy and sense of belonging. This group included students in fifth and seventh grades who participated in the math workshop intervention and those in sixth and eighth grades who did not[4]. The results in this research study include only those students (n=203) who gave assent and whose parents provided consent for their participation in this research study. The research sample represents approximately 70% of the student body in the middle school.

Table 27

Student Respondents on Pre/Post Math Attitude Survey 2016 to 2017

Grade	Gender		Total
	Girls	Boys	
Fifth	26	25	51
Sixth	32	34	66
Seventh	26	31	57
Eighth	11	18	29
Total	95	108	203

A mean score for each of the two constructs, self-efficacy and sense of belonging, was calculated for each student. As anticipated based on the needs assessment, the mean scores on these constructs for girls in the middle school were lower than the mean scores for boys. The difference in mean scores for self-efficacy in the fall was statistically significant by gender,

[4] Students in these grades did, however, participate in the pilot year of the program, making them less ideal as comparison group.

t(201)=2.06, p=.041, d=.30. The difference in mean sense of belonging scores was also statistically significant, t(201)=1.99, p=.048, d=.28.

There were a number of patterns in the fall survey data that are helpful to recognize (Table 28). First, in every grade boys scored higher on both constructs compared to girls. Second, girls' self-efficacy scores appeared to decline during middle school, whereas boys' scores stay relatively stable. There also appear to be a dip in sense of belonging scores for older girls, although it is not as dramatic as the change in self-efficacy. Thus, although it is beyond the scope of this study, it may be worthwhile in future research to consider in more detail the developmental trajectory of self-efficacy and sense of beloning in middle-school girls and to assess an intervention against a projected change based on gender and age.

Table 28

Mean Self-Efficacy and Belonging Scores by Grade and Gender September 2016 (on a Scale From 0 to 4)

Grade	Mean Self-Efficacy		Mean Sense of Belonging	
	Girls	Boys	Girls	Boys
Fifth	3.01	3.07	3.33	3.38
Sixth	2.95	3.17	3.21	3.45
Seventh	2.85	2.95	3.25	3.30
Eighth	2.66	3.12	3.20	3.35

Fall 2016 Survey Correlations with Mathematics Achievement

Fall survey results for students' mean self-efficacy and sense of belonging scores were correlated with mathematics achievement scores from 2016. The purpose of this statistical test was to determine if there was a meaningful relationship between students' self-reported attitudes

on the survey and mathematics achievement prior to the intervention. This analysis revealed a statistically significant small to medium linear relationship between middle school students' self-efficacy and mathematics achievement scores and students' sense of belonging and their mathematics achievement scores.

The same analysis was then repeated selecting first for boys and then for girls. The results of this analysis were quite different and revealed that while boys' self-efficacy and sense of belonging were moderately correlated with mathematics achievement, girls' scores on these two constructs were not correlated with mathematics achievement (Tables 29 and 30). The same result was found with spring survey responses and spring mathematics achievement test scores. It is not simple to interpret this result, as it contradicts some of the literature and needs assessment data suggesting these two constructs are highly associated with mathematics achievement. However, the needs assessment did not break down the correlations to see if they were different by gender. It is possible that girls did not report their experience as accurately or honestly as boys, or it may be that girls' attitudes about mathematics are not as associated with achievement as those of boys. This will be a potentially significant question to investigate further in future applied educational research.

Table 29

Fall 2016 Correlations Between Middle School Boys' Self-Efficacy, Sense of Belonging and Mathematics Achievement Scores

	Mean Belonging Fall	Mean Self-Efficacy Fall	Mathematics Scaled Score 2016
Mean belonging fall Pearson Correlation Sig. (2-tailed)	1	.455** .000	.281* .019
Mean self-efficacy fall Pearson correlation Sig (2-tailed)	.455** .000	1	.443** .000
Math scale score 2016 Pearson Correlation Sig (2-tailed)	.281** .019	.443** .000	1

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlations is significant at the 0.05 level (2-tailed).
c. Listwise N = 69

Table 30

Fall 2016 Correlations Between Middle School Girls' Self-Efficacy, Sense of Belonging and Mathematics Achievement Scores

	Mean Belonging Fall	Mean Self-Efficacy Fall	Mathematics Scaled Score 2016
Mean belonging fall Pearson Correlation Sig. (2-tailed)	1	.625**	.085
Mean self-efficacy fall Pearson correlation Sig (2-tailed)	.625** .000	1	.067 .555
Math scale score 2016 Pearson Correlation Sig (2-tailed)	.085 .450	.067 .555	1

**. Correlation is significant at the 0.01 level (2-tailed).
b. Listwise N = 81

Spring Survey Results

In May 2017, 260 middle school students in each grade re-took the mathematics attitude survey (91% participation), of which 240 also had scores from September 2016. Of this group, the data from 203 students who provided parental consent were reported in this study. For each student, their change in self-efficacy and sense of belonging was computed by subtracting their mean fall self-efficacy and sense of belonging scores from their mean spring scores (Table 29).

Table 31

A Comparison of Student Mean Self-Efficacy and Belonging Scores by Grade and Gender September 2016 and May 2017

Grade	Mean Self-Efficacy				Mean Sense of Belonging			
	Girls		Boys		Girls		Boys	
	Fall	*Spring*	*Fall*	*Spring*	*Fall*	*Spring*	*Fall*	*Spring*
Fifth	3.01	2.96	3.07	3.16	3.33	3.42	3.38	3.53
Sixth	2.95	2.84	3.17	3.26	3.21	3.19	3.45	3.31
Seventh	2.85	2.93	2.95	2.99	3.25	3.34	3.30	3.38
Eighth	2.66	2.68	3.12	3.01	3.20	3.14	3.35	3.23

Note. Blue=score increase, Red=score decrease.

The results of a 3-way ANOVA for the main effects of gender (boy, girl), track (on-level, advanced) and intervention (intervention, control) on student's sense of belonging in mathematics found a main effect for the intervention, $F(1)=6.610$, $p=.011$, $\eta 2 = .040$. In other words, there was an observable difference in self-reported belonging between students in both fifth and seventh grades who received the math workshop intervention and those that did not. Students in fifth and seventh grades, both boys and girls, on average saw a small increase in their sense of belonging score while students in sixth and eighth grades had a drop in reported sense of belonging. Furthermore, this difference between students who received math workshop and

students who did not receive math workshop was statistically significant, $t(200)=2.299$, $p=.023$. However, the research question of this study specifically sought to determine if girls saw a greater increase in sense of belonging than boys. The results indicate of the 3-way ANOVA found no statistically significant effect for gender on sense of belonging, $F(1)=.514$, $p=.475$, $\eta 2 =.003$.

A 3-way ANOVA for the main effects of gender (boy, girl), track (on-level, advanced) and intervention (intervention, control) on students' self-efficacy found no statistically significant effects.

Research Question 2

Did a change in girls' self-efficacy correlate with a change in mathematics achievement test scores? For fifth grade girls, there was a moderate *negative* linear relationship between the change in students' reported self-efficacy scores and the change in mathematics achievement scores, $r(23)= -.429$, $p=.041$. This was an unexpected finding. For fifth grade girls, those who reported an increase in their self-efficacy in mathematics were more likely to see a reduction in mathematics achievement scores and those girls who reported a decrease in self-efficacy were more likely to have an increase in test-scores. One potential explanation is that "tracking" begins in fifth grade and stronger female students exposed to an accelerated mathematic curriculum that might teach more of the skills on the achievement test would also be vulnerable to a drop in self-efficacy and sense of belonging. The advanced fifth grade mathematics class has a male to female ratio of 2:1 and this experience with ability grouping could be one source of stereotype threat for fifth grade girls. The data support this hypothesis because a change in self-efficacy was moderately negatively correlated with a change in mathematics achievement for on-level fifth grade girls, but was more strongly negatively correlated for advanced fifth grade girls.

Table 32

Correlation Between Fifth Grade Girls' Change in Mathematics Scaled Scores and Self-Efficacy 2016-2017

Track1617			Change in Mathematics Scaled Score 2016-2017	Change in Self-Efficacy
On-level	Change in Mathematics Scaled Score 2016-2017	Pearson Correlation	1	-.253
		Sig. (2-tailed)		.404
		N	13	13
	Change in Self-Efficacy	Pearson Correlation	-.253	1
		Sig. (2-tailed)	.404	
		N	13	16
Advanced	Change in Mathematics Scaled Score 2016-2017	Pearson Correlation	1	-.475
		Sig. (2-tailed)		.101
		N	13	13
	Change in Self-Efficacy	Pearson Correlation	-.475	1
		Sig. (2-tailed)	.101	
		N	13	15

In seventh grade, there was no statistically significant correlation between change in girls' self-efficacy and change in mathematics achievement score.

Table 33

Correlations Between Seventh Grade Girls Change in Mathematics Scaled Scores and Change in Self-Efficacy 2016-2017

		Change in Mathematics Scaled Score 2016-2017	Change in Self-Efficacy
Change in Mathematics Scaled Score 2016-2017	Pearson Correlation	1	.247
	Sig. (2-tailed)		.196
	N	29	29
Change in Self-Efficacy	Pearson Correlation	.247	1
	Sig. (2-tailed)	.196	
	N	29	29

Research Question 3

Did a change in girls' sense of belonging correlate with a change in mathematics achievement test scores? For the middle school as a whole (fifth through seventh grades), there was a statistically significant positive correlation between change in sense of belonging and change in mathematics achievement. However, this relationship did not remain when cohorts were broken down by gender and grade.

Table 34

Correlations for Sense of Belonging and Change in Mathematics Achievement for Middle School Students 2016-2017

		Change in Belonging	Change in Mathematics Scaled Score 2016-2017
Change in Belonging	Pearson Correlation	1	.182*
	Sig. (2-tailed)		.028
	N	203	146
Change in Mathematics Scaled Score 2016-2017	Pearson Correlation	.182*	1
	Sig. (2-tailed)	.028	
	N	146	146

*. Correlation is significant at the 0.05 level (2-tailed).

There was no statistically significant correlation between fifth or seventh grade girls' change in self-reported sense of belonging and mathematics achievement test score.

Table 35

Correlations Between Fifth Grade Girls' Change in Mathematics Scaled Score and Sense of Belonging 2016-2017

		Change in Mathematics Scaled Score 2016-2017	Change in Belonging
Change in Mathematics Scaled Score 2016-2017	Pearson Correlation	1	.119
	Sig. (2-tailed)		.587
	N	23	23
Change in Belonging	Pearson Correlation	.119	1
	Sig. (2-tailed)	.587	
	N	23	26

Table 36

Correlations Between Seventh Grade Girls' Change in Mathematics Scaled Score and Sense of Belonging 2016-2017

		Change in Mathematics Scaled Score 2016-2017	Change in Belonging
Change in Mathematics Scaled Score 2016-2017	Pearson Correlation	1	.110
	Sig. (2-tailed)		.568
	N	29	29
Change in Belonging	Pearson Correlation	.110	1
	Sig. (2-tailed)	.568	
	N	29	29

Interview Results

In this research study, the interview results were used to help understand and explain the patterns identified in the quantitative data. As discussed in Chapter 4, twelve girls were randomly selected for brief semi-structured interviews about their experience in math workshop, three students from each of the four sections of girls. One selected student declined to participate and was replaced with another randomly chosen participant who did agree to an interview. The Director of Research at the BC School individually interviewed all of the participating students during study hall for approximately 10 minutes. Analysis of interview data suggested a number of factors such as peer support and engaging open-ended tasks that may protect against the negative effects of stereotype threat as well as a number of factors that may increase vulnerability to

stereotype threat such as high-stakes assessments, competition, and fear of social judgment regarding mathematics ability.

Sources of Self-Efficacy and Belonging

A particularly robust theme in the interviews was the idea that working with other students in a collaborative environment tended to increase both girls' self-efficacy and sense of belonging. Eight of the students interviewed mentioned working with other students as a source of comfort and confidence both in their regular mathematics classes and in math workshop. Fifth grade girls were more likely to use the term "friends" explicitly with four of the six fifth grade girls referred to being with friends when asked what helps them feel comfortable in mathematics. For example, a fifth grade girl said:

> What make me feel comfortable is when my friends are one sitting next to me because sometimes teachers are like oh we can't sit friends next to friends because they'll talk all the time but I know that I'm not going to talk because they help me and I know that I trust them to give me help not just the right answers because that's what teachers assume friends are doing but honestly we're not but I love it when my friends are there and I have the supplies I need. (CT 5, May 2017).

In this quotation, the student describes her friend as a source of potential assistance and also in a similar manner to the "supplies" that she needs. It is as if her friends are an important tool for her in tackling a mathematics problem. Seventh grade girls also talked about the importance of peers and working with other people as a source of help and encouragement although they were less likely to use the term "friends." One of the six seventh grade students used the term friends, while 4 of the 6 fifth grade students did.

Of the twelve students interviewed, five explicitly talked about the single-gender nature of the classroom, and four of these five students were in fifth grade. All five students mentioned the all-girls grouping as a positive aspect of the class that increased self-efficacy. For example, in the response to the question: "What do you see as different about math workshop compared to your regular math class," a fifth grade girl articulated:

> I like how it's only one gender and it's just girls. It makes me feel stronger and sometimes like I can feel if there is...sometimes in my regular math class there can be some boys that always call out the answers and don't like let everyone else just think about it. So it makes me feel better to have a group of girls working together and it makes me feel stronger. And no one calls out most of the time just they usually just they got the answer and they all work together and that makes me feel good. (JPN5, May 2017)

In this case, the student reports high self-efficacy as feeling "stronger" because she knows that she will have time to "think about it" and that in her group they will "work together." Another fifth grade girl spoke directly about the negative stereotype about girls in mathematics and how she feels that an all-girls learning environment was beneficial:

> other girls that sometimes are more timid in regular math class come out more with their answers and with their like actual solving of the problems and I see sometimes in regular math class they don't solve the problems at all because they know that they'll just get them wrong which I think is totally not true that girls totally have the ability to do that and stuff but I think they come out more in math workshop which is good but I think they need to learn to get it in math, regular math (CT 5, May 2017)

In this student's words, girls in math workshop were more likely to "come out more with their answers" in math workshop and be less timid. Another fifth grade student stated simply, "Math

Workshop is all girls. I guess I feel more comfortable saying my answer (CR5, June 2017). Another seventh grade student commented that she preferred the all-girls environment because, "Well I feel like we have more in common so yeah. Like I look forward going to the class." (AB 7, May 2017). Students in both fifth and seventh grades put a high value on working with peers that they trusted and reported both higher self-efficacy and sense of belonging when they worked with a partner or partners.

Another theme in the interviews was the idea that math workshop was more fun and enjoyable than regular mathematics class. Of the 12 students interviewed, nine of them used the term "fun" when talking about math workshop. When asked "What was different about how math workshop compared to your regular mathematics class, one seventh grade student said, "Well I feel like math workshop is more of a fun math class. And like it's really fun because there are no other class where it is all girls in one class and all boys in another" (AB 7, May 2017). In this case, the student identified that the gender division was a source of enjoyment for her.

In other interviews, students mentioned the material of the class as a source of fun. One fifth grade girl differentiated math workshop from regular math class saying, "Math workshop is a littler harder so it's more challenging and it's a little bit more fun" (CR5, May 2017), while another said "it trains your mind in the same way math does but not just a repeat of math class which is very fun. I like it" (CT5, May 2017). Seventh grade students also mentioned this theme. One girl described how math workshop was different saying:

> The things we do are kind of more practical. I don't think that's the right word to use but like they are more fun. We like experiment into stuff. We do more proofs. Proof is the right word. And we work in groups instead of individual (EB 7, May 2017).

The majority of the students mentioned math workshop as being more pleasurable than their regular math class. The reasons they gave included the gender grouping, the chance to work with friends and the curriculum that permitted more exploration and discovery.

Sources Inhibiting Self-Efficacy and Belonging

Although friends and peers were frequently mentioned as a source of comfort, students also reported that social dynamics could be a source of potential discomfort as well. Several students talked about feeling badly when they were comparing themselves with peers in some manner or felt that other students were judging them. A fifth grade student said she felt less comfortable in the math workshop setting because, "I'm so used to my normal math class and there's some girls in there [math workshop] that like I don't know" (JS5, May 2017). This sensitivity to partners was common with both grade levels. Seventh grade students also reported some anxiety about working with certain peers. One seventh grade student described, "A lot of times when I'm partnered with someone I'm not like with anyone who I'm friends with it's a little bit like awkward (ER7, May 2017). It became clear through the interviews that the "right" partner or partners was a primary concern of the students in both grades.

There was also some degree of awareness in of the difference in students' mathematical knowledge and experience, and two students mentioned this as a source of some frustration. Fifth grade is the first year that students are divided by prior achievement and grouped into "on-level" and "advanced" mathematics classes. By seventh grade, an additional group is created so that there are three divisions, on-level pre-algebra, advanced pre-algebra and Algebra I. Students in both fifth and seventh grade reported awareness of these distinctions. For example, a fifth grade student in the on-level mathematics class suggested students be grouped both by gender and by "level" because as she explains:

> some people finish more because they're on another level, like a higher level so they can do better things and like if we're doing a puzzle and they're like let's find the median or average or something and I was like what's that and then they have to explain and stuff (CR5, June 2017)

Although this situation could also be viewed as a positive way to learn from a peer, at least some of the students found working with peers from different mathematics classes to be challenging. A seventh grade girl in the "on-level" mathematics section, the least advanced of the three, articulated a desire for more independent work:

> sometimes my partner or my group kind of knows more how to do the problem or how to figure it out and you know they kind of do it or they're like okay this is how you do it and then I'm like okay and I just you know it doesn't push me to kind of think on my own… (FL7, May 2017).

Both students who mentioned this source of discomfort were in the "on-level" mathematics section. A student in the most advanced seventh grade mathematics section also raised the issue of different levels of mathematics, but the mixed-level grouping did not bother her, "I don't have a lot of the people in the math workshop class in the mathematics class. There are a bunch of different levels but there's also like stuff for everyone to do." (EB7, May 2017). These comments suggest that it is important to remember that students come to the math workshop with a keen awareness and sensitivity to the level of their "regular" math class and that it is necessary to find activities that allow for all students to feel successful at mathematical problem solving while also being provided with appropriate challenges.

Students reported feeling most uncomfortable and discouraged when they felt pressured to give an answer and they feared they would be judged by peers. One fifth grade student described an uncomfortable moment in mathematics class:

> Sometimes when I do hesitate to give an answer maybe because I lost my page in my math book or some other reason. People start looking at me like come on the next question I have the answer too and it's kind of scary because then it makes me even more like baffled that I can't do it so I just want to because I wish I had like an extra minute to just breathe before I can answer before people start like looking at me like umm what are you doing? Answer the question (Inaudible) and I have the answer but yeah sometimes that makes me a little off (CT5, May 2016).

This vivid description evokes a high degree of anxiety about the "performance" nature of mathematics in which students are called on to answer independently a question within a time limit. Several students mentioned that participating in class made them feel nervous. This theme was also present in the seventh grade. One student reported she felt uncomfortable when, "When I have to raise my hand. I mean like when I have to do the problems on the board and when I like am talking in class because I might have the answer wrong" (EB7, May 2016).

Finally, six of the twelve respondents mentioned tests and quizzes as times that they feel uncomfortable or nervous. Several students explicitly mentioned the fact that tests and quizzes can raise feelings of nervousness because "we are all independent" (RK7, May 2016) and there is no opportunity for collaboration. The feedback from the interviewees suggests that some girls have a more positive experience when they are given more time to think through an answer and respond either in a group or prepare an answer in advance to present to the class.

Conclusion

This mixed methods study has a number of key findings that can further the conversation about how to close the gender gap between high ability boys and girls regarding both their self-efficacy and sense of belonging in mathematics as well as their success on standardized tests of achievement. The following results were the most salient and supported by multiple forms for data:

- The math workshop may have helped to reduce the gender gap in achievement between advanced students, but the results were more mixed for on-level students
- The math workshop intervention improved students' self-reported sense of belonging in mathematics for the majority of participating students, both boys and girls
- The efficacy of the intervention may have been moderated by students' increased sense of belonging due to the "friend effect," the opportunity to work in pairs or groups with trusted peers, which students reported finding fun and less stressful than timed, individual work

Discussion

Practitioners in coeducational environments, and particularly those working with high-ability students, may want to consider offering an all-girls' option for mathematics instruction. The results from this intervention suggest that even an occasional supplementary class may be helpful in allowing girls an opportunity for practicing mathematics in which stereotype threat is reduced and may help to close achievement gaps on standardized tests. Furthermore, girls tended to respond positively to a class that was inquiry based and collaborative, and this model of learning may be less likely to create competitive or evaluative situations that heighten stereotype threat. There did not appear to be any negative effects on boys' attitudes or achievement in

mathematics, and their survey responses indicated that that the intervention may have helped improve their sense of belonging in mathematics.

When introducing the program in the pilot year, middle school students had difficulty understanding the concepts of stereotype threat, and excessive discussion about this topic may have heightened gender awareness instead of lessening it as intended. Although prior research by Johns et al. (2005) suggested that educating girls about stereotype threat could reduce its negative impact, this may not hold true for younger students. The research conducted by Johns et al. (2005) used college-age participants who would be better equipped to understand the complexities of stereotype threat. In response to the Pilot Year experience in the second year, the emphasis was on educating parents and teachers about the program goals and creating a fun and collaborative classroom environment for the students. This appeared to improve the response to the program. When adults in the community are fully informed about the reason for the program and the activities are engaging for all learners, there may be less resistance from participating students. Furthermore, the program was adjusted by allowing students to select their own gender identity. This reasonably small administrative change was intended to give students a greater sense of choice over their assignment to either a group of girls or boys. If a students identified their gender as non-binary, they would be permitted to work separately or join either group. Given the ongoing conversations within this school's context and nationally about the spectrum of gender identity, this may become an increasingly important issue for this type of intervention.

Limitations and Future Research Directions

This research study had a number of limitations including sample size and the absence of a comparison group. All participants in the study attended the same school and come from a similarly affluent socioeconomic and culture background that is majority white with high

achievement in both verbal reasoning and mathematics. Therefore, it may be difficult to generalize from this sample to other populations. The sample size was also relatively small due to the challenge of obtaining permission from parents and students to use both survey and standardized test data in this study. The study sample may represent students who are more comfortable in mathematics and therefore more willing to share their personal data regarding mathematics.

Without a randomized control group, it is difficult to assess if the intervention was more helpful than it appears because girls' scores on these constructs might have declined without intervention and/or boys' scores may have risen without the intervention. The challenges in interpreting these results underscores the necessity of a randomized controlled study in which the effects of age and gender could be controlled. Many educational research designs lack a control group for the studied population because students cannot be randomly assigned to a single-gender or coeducational setting (Cherney & Campbell, 2011). Therefore, results in these cases cannot be considered causal. Although there are a few examples of randomized studies investigating all-girls classes (e.g., Eisenkopf et al., 2014; Kessels & Hannover, 2008) a large-scale randomized experiment in which students were assigned to either a mixed-gender or single-gender classroom for a male-stereotyped course in the United States would contribute to understanding when and if girls benefit from all-girl educational settings in school. The results of this study suggest that students' levels of self-efficacy and sense of belonging fluctuate as they mature. Therefore, it would also be advisable to conduct a longitudinal study to understand better how student social-emotional and academic maturation interact with attitudes about mathematics.

Another limitation common to studies of the relationship between mathematics self-efficacy and achievement is the reliance on self-report measures of self-efficacy (Kessels & Hannover, 2008; Preckel et. al, 2008). Many students are not aware of their own feelings about mathematics, and may additionally be unmotivated to complete a survey thoughtfully. Given that stereotype threat functions at a largely implicit level, it may be difficult to measure changes in self-efficacy and sense of belonging with an explicit report measure. Future studies may consider using a test such as the Implicit Attitudes Test (IAT), which aims to assess attitudes that exist below conscious awareness. It would also be helpful to have more qualitative research in this research area to understand better the student experience in single-sex classrooms. This study was limited by both time and resources, but future studies should consider interviewing both boys and girls to garner greater insight into students' lived experiences.

Finally, this study altered both the gender grouping of the mathematics classes as well as the pedagogy of the class. It is possible that some of benefit to girls' achievement scores resulted from experiencing a more inquiry-focused approach to mathematics and was not related to the gender grouping. In future studies, it will be important to control for content delivery in order to better determine if and when an all-girls mathematics class is beneficial.

References

American Psychological Association. (2011). The guidelines for psychological practice with lesbian, gay, and bisexual clients, adopted by the APA Council of Representatives. Washington, DC: APA.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52*(1), 1–26. doi: https://doi.org/10.1146/annurev.psych.52.1.1

Barreto, M., & Ellemers, N. (2005). The perils of political correctness: Men's and women's responses to old-fashioned and modern sexist views. *Social Psychology Quarterly*, *68*(1), 75–88. doi: https://doi.org/10.1177/019027250506800106

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science*, *16*, 101–105. doi: 10.1111/j.0956-7976.2005.00789.x

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' mathematics anxiety affects females' mathematics achievement. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 1860–1863. doi: 10.1073/0910967107

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*, 389-391.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101.

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, *106*, 676.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*, 75–141. doi: 10.1177/1529100614541236

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, *135*, 218. doi: 10.1037/a0014412

Chernail, R. J. (2011). Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research. *The Qualitative Report*, *16*, 255–262.

Cherney, I. D., & Campbell, K. L. (2011). A league of their own: Do single-sex schools increase females' participation in the physical sciences? *Sex Roles*, *65*, 712–724. doi: 10.1007/s11199-011-0013-6

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2016). Why are some STEM fields more gender balanced than others? *Psychological Bulletin, 143,* 1-35. http://dx.doi.org/10.1037/bul0000052

Clune, D. (2014, October 9). *CTP mathematics CCSSM alignment study*. New York, NY: Educational Records Bureau. Retrieved from https://www.erblearn.org/news/ctp-mathematics-ccssm-alignment-study

Creswell, J. W., & Clark, V. L. P. (2007). *Designing and conducting mixed methods research*. Los Angeles, CA: Sage.

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*, 766–779. doi: 10.1111/j.1467-8624.2010.01529.x

Dasgupta, N. (2011). Ingroup experts and peers as social vaccines who inoculate the self-concept: The stereotype inoculation model. *Psychological Inquiry*, *22*, 231–246. doi: 10.1080/1047840X.2011.607313

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). *The effects of Teach for America on students: Findings from a national evaluation* (pp. 8792–750). Madison, WI: University of Wisconsin-Madison, Institute for Research on Poverty.

Doris, A., O'Neill, D., & Sweetman, O. (2013). Gender, single-sex schooling and maths achievement. *Economics of Education Review*, *35*, 104–119.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Taylor & Francis.

Dusenbury, L. Brannigan, R., Falco, M & Hansen, W.B.(2003). A review of research on fidelity of implementation implications for drug abuse prevention in school settings. *Health Education Research*, *18*, 237-256.

Educational Records Bureau. (2014a). Comprehensive Testing Program Online: Technical Report. Retrieved from: http://staging.programworkshop.com/6.0.0.0/custom/erb/CTP%20Online%20Technical%20Report.pdf

Educational Records Bureau (2014b). Comprehensive Testing Program Growth Report Modifications. Retrieved from: https://www.programworkshop.com/PW2/Core/2.0/Login/Login/LoginSuccess?skin=default&programid=80&pw2=1&sk=67&cv=2.0&rv=1.1&scv=3.1&sbv=1.1&sc=I005B0D644038467E154D356424AC173D2147D7E90368

Educational Records Bureau. (2017). Spring Norms Book 2017. Retrieved from: https://www.programworkshop.com/6.0.0.0/admin/configuration/loadpreview.aspx?programid=80&permissionid=647&filename=2017_Spring_ERB_Norms_Book.pdf&permission=Custom_80_647&sc=I007319930596167C1D870A94157B0D85108B287B9F13

Egorova, M. S. (2016). Sex differences in mathematical achievement: grades, national test, and self-confidence. *Psychology in Russia*, *9*(3), 4.

Eisenkopf, G., Hessami, Z., Fischbacher, U., & Ursprung, H. W. (2014). Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland. *Journal of Economic Behavior & Organization*, *115*, 123–143. doi: http://dx.doi.org/10.1016/j.jebo.2014.08.004

Eliot, L. (2013). Single-sex education and the brain. *Sex Roles*, *69*, 363–381. doi: 10.1007/s11199-011-0037-y

Ellison, G., & Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the American Mathematics Competitions. *Journal of Economic Perspectives*. 24, 109–128. doi: 10.1257/jep.24.2.109

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*, 103-127. doi: 10.1037/a0018053

Field Medal Details (2014, May fifth) Retrieved from http://www.mathunion.org/general/prizes/fields/details/

Feniger, Y. (2011). The gender gap in advanced math and science course taking: does same-sex education make a difference? *Sex Roles*, *65*, 670–679. doi: 10.1007/s11199-010-9851-x

Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of Mathematics Interest in Adolescence: Influences of Gender, Family and School Context. *Journal of Research on Adolescence*, 20, 507-537. doi: 10.1111/j.1532-7795.2010.00645.x

Fryer Jr, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*, 210–240. doi: 10.1257/app.2.2.210

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, *77*, 337–353. doi: 10.1006/jecp.2000.2594

Glesne, C., & Peshkin, A. (2010). *Becoming qualitative researchers: An introduction* (4th ed.). White Plains, NY: Longman.

Goldman, S., & Booker, A. (2009). Making mathematics a definition of the situation: Families as sites for mathematical practices. *Anthropology & Education Quarterly*, *40*, 369–387. doi: 10.1111/j.1548-1492.2009.01057.x

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, *102*, 700–717. doi: 10.1037/a0026659

Green, J., Willis, K., Hughes, E., Small, R., Welch, N., Gibbs, L., & Daly, J. (2007). Generating best evidence from qualitative research: the role of data analysis. *Australian and New*

Zealand Journal of Public Health, *31*(6), 545–550. doi: 10.1111/j.1753-6405.2007.00141.x

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science,* New series, *320*(5880), 1164–1165. doi: 10.1126/science.1154094

Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related mathematics attitudes. *Sex Roles*, *66*, 153–166.

Hall, J. M., & Ponton, M. K. (2005). Mathematics self-efficacy of college freshman. *Journal of Developmental Education, 28*(3), 26-30.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1–51. doi: https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hargreaves, M., Homer, M., & Swinnerton, B. (2008). A comparison of performance and attitudes in mathematics amongst the 'gifted'. Are males better at mathematics or do they just think they are? *Assessment in Education: Principles, Policy & Practice*, *15*, 19–38. doi: 10.1080/09695940701876037

Hausmann, R. & Tyson, L. (2015). *The global gender gap report 2015*. Geneva, Switzerland: World Economic Forum. Retrieved from: https://www.weforum.org/reports/global-gender-gap-report-2015/

Herbert, J., & Stipek, D. (2005). The emergence of gender differences in children's perceptions of their academic competence. *Journal of Applied Developmental Psychology*, *26*, 276–295. doi: https://doi.org/10.1016/j.appdev.2005.02.007

Hilliard, L. J., & Liben, L. S. (2010). Differing levels of gender salience in preschool classrooms: Effects on children's gender attitudes and intergroup bias. *Child Development*, *81*, 1787–1798.

Hoffman, B. H., Badgett, B. A., & Parker, R. P. (2008). The effect of single-sex instruction in a large, urban, at-risk high school. *The Journal of Educational Research*, *102*, 15–36.

Hubbard, L., & Datnow, A. (2005). Do single-sex schools improve the education of low-income and minority students? An investigation of California's public single-gender academies. *Anthropology & Education Quarterly*, 36,115–131. doi: 10.1525/aeq.2005.36.2.115

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, *11*, 365–371. doi: https://doi.org/10.1111/1467-9280.00272

Inzlicht, M., & Ben-Zeev, T. (2003). Do high-achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology*, *95*, 796-805. doi: 10.1037/0022-0663.95.4.796

Jansen, P., Zayed, K., & Osmann, R. (2016). Gender differences in mental rotation in Oman and Germany. *Learning and Individual Differences*, *51*, 284–290. doi: https://doi.org/10.1016/j.lindif.2016.08.033

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle teaching stereotype threat as a means of improving women's math performance. *Psychological Science*, *16*, 175–179. doi: 10.1111/j.0956-7976.2005.00799.x

Kane, J. M., & Mertz, J. E. (2012). Debunking myths about gender and mathematics performance. *Notices of the AMS*, *59*, 10–21.

Kessels, U., & Hannover, B. (2008). When being a girl matters less: Accessibility of gender-related self-knowledge in single-sex and coeducational classes and its impact on students' physics-related self-concept of ability. *British Journal of Educational Psychology*, *78*, 273–289. doi: 10.1348/000709907X215938

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes a prospective study of female college students. *Psychological Science*, *18*, 13–18. doi: https://doi.org/10.1111/j.1467-9280.2007.01841.x

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*, 1123–1135. doi:10.1037/a0021276

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M. W., … Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: Institute of Education Services National Center for Special Education Research. Retrieved from http://ies.ed.gov/ncser/pubs/20133000/

Mathematical Association of America. (2016). *AMC historical statistics*. Washington, DC: Mathematical Association of America. Retrieved from http://amc-reg.maa.org/reports/generalreports.aspx

Mael, F., Alonso, A., Gibson, D., Rogers, K., & Smith, M. (2005). *Single-sex versus coeducational schooling: A systematic review*. Doc# 2005-01. Washington, DC: US Department of Education. Retrieved from http://www.ed.gov/about/offices/list/opepd/reports.html

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin*, *28*, 1183–1193. doi: https://doi.org/10.1177/01461672022812004

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat how situational cues affect women in math, science, and engineering settings. *Psychological Science*, *18*, 879–885. doi: 10.1111/j.1467-9280.2007.01995.x

National Center for Education Statistics. (2014). *Digest of Education Statistics*. Washington, DC: U.S. Department of Education. Retrieved from https://nces.ed.gov/programs/digest/2014menu_tables.asp

Neuville, E., & Croizet, J. C. (2007). Can salience of gender identity impair math performance among 7–8 years old girls? The moderating role of task difficulty. *European Journal of Psychology of Education*, *22*(3), 307–316. doi: https://doi.org/10.1007/BF03173428

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *The Journal of Economic Perspectives*, *24*, 129–144. doi: 10.1257/jep.24.2.129

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … Kesebir, S. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*, 10593–10597.

Organisation for Economic Co-operation and Development (OECD). (2014). *PISA 2012 Results: Creative problem solving: Students' skills in tackling real-life problems (Volume V)*. Paris, France: OECD Publishing.

OECD. (2015). *The ABC of gender equality in education: Attitude, behaviour and confidence.* Paris, France: OECD Publishing. doi: 10.1787/9789264229945-en

Pahlke, E., Hyde, J. S., & Allison, C. M. (2014). The effects of single-sex compared with coeducational schooling on students' performance and attitudes: A meta-analysis. *Psychological Bulletin, 140*, 1042-1072.

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, *86*, 193–203. doi: 10.1037/0022-0663.86.2.193

Park, H., Behrman, J. R., & Choi, J. (2013). Causal effects of single-sex schools on college entrance exams and college attendance: Random assignment in Seoul high schools. *Demography*, *50*(2), 447–469. doi: 10.1007/s13524-012-0157-1

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, *37*, 239–253. doi: 10.1016/j.ssresearch.2007.06.012

Picho, K., & Stephens, J.M.(2012). Culture, Context and Stereotype Threat: A Comparative Analysis of Young Ugandan Women in Coed and Single-Sex Schools. The Journal of Educational Research, 105, 52-63. doi:10.1080/00220671.2010.517576

Poggenpoel, M., & Myburgh, C. (2003). The researcher as research instrument in educational research: A possible threat to trustworthiness? *Education*, *124*, 418-423.

Pope, D. G., & Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *The Journal of Economic Perspectives*, 24, 95-108. doi:10.1257/jep.24.2.95

Preckel, F., Goetz, T., Pekrun, R., & Kleine, M. (2008). Gender differences in gifted and average-ability students comparing females' and males' achievement, self-concept, interest, and motivation in mathematics. *Gifted Child Quarterly*, *52*, 146–159. doi: https://doi.org/10.1177/0016986208315834

Quaiser-Pohl, C., Jansen, P., Lehmann, J., & Kudielka, B. M. (2016). Is there a relationship between the performance in a chronometric mental-rotations test and salivary testosterone and estradiol levels in children aged 9–14 years? *Developmental Psychobiology*, *58*, 120–128.

Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2013). Mathematics anxiety, working memory, and mathematics achievement in early elementary school. *Journal of Cognition and Development*, 14, 187–202. doi: 10.1080/15248372.2012.664593

Rattan, A., Good, C., & Dweck, C. S. (2012). "It's ok—Not everyone can be good at math": Instructors with an entity theory comfort (and demotivate) students. *Journal of Experimental Social Psychology*, *48*, 731–737. doi: 10.1016/j.jesp.2011.12.012

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*, 268–302. doi: https://doi.org/10.3102/0002831210372249

Rogers, E.M., (2003). *Diffusion of innovations* (Fifth Edition). New York, NY: Free Press.

Rossi, P., Lipsey, M., & Freeman, H. (2004). *Evaluation: A systematic approach.* Thousand Oaks, CA: Sage.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics*, *102*, 245–253. doi: 10.1111/j.1949-8594.2002.tb17883.x

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*, 440-452. doi: 10.1037/0022-3514.85.3.440

Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles*, *50*, 835–850. doi: https://doi.org/10.1023/B:SERS.0000029101.74557.a0

Schneeweis, N., & Zweimüller, M. (2012). Females, females, females: Gender composition and female school choice. *Economics of Education Review*, *31*, 482–500. doi: https://doi.org/10.1016/j.econedurev.2011.11.002

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Signorella, M. L., Hayes, A. R., & Li, Y. (2013). A meta-analytic critique of Mael et al.'s (2005) review of single-sex schooling. *Sex Roles, 69*(7-8), 423-441. doi:10.1007/s11199-013-0288-x

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, *42*, 70-83. doi: 10.1037/0012-1649.42.1.70

Smith, J. L., Lewis, K. L., Hawthorne, L., & Hodges, S. D. (2013). When trying hard isn't natural: Women's belonging with and motivation for male-dominated stem fields as a function of effort expenditure concerns. *Personality and Social Psychology Bulletin*, *39*, 131–143. doi: 10.1177/0146167212468332

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*, 4–28. doi: https://doi.org/10.1006/jesp.1998.1373

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613. doi: http://dx.doi.org/10.1037/0003-066X.52.6.613

Steele, C. (2011). *Whistling Vivaldi: And other clues to how stereotypes affect us (Issues of our time)*. New York, NY: WW Norton & Company.

Steele, J. R., & Ambady, N. (2006). "Math is Hard!" The effect of gender priming on women's attitudes. *Journal of Experimental Social Psychology*, *42*, 428–436. doi: https://doi.org/10.1016/j.jesp.2005.06.003

Stoet, G., Bailey, D. H., Moore, A. M., & Geary, D. C. (2016). Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. *PloS One*, *11*(4), e0153857.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher, 36*, 187–198. doi: 10.3102/0013189X07303396

Tapia, M. (1996). The Attitudes toward Mathematics instrument. *Paper presented at the Annual Meeting of the Mid-South Educational Research Association*. Tuscaloosa, AL.

Tully, D., & Jacobs, B. (2010). Effects of single-gender mathematics classrooms on self-perception of mathematical ability and post secondary engineering paths: An Australian case study. *European Journal of Engineering Education*, *35*, 455–467. doi: http://dx.doi.org/10.1080/03043797.2010.489940

Turner III, D. W. (2010). Qualitative interview design: A practical guide for novice investigators. *The Qualitative Report*, *15*(3), 754-760.

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, *34*, 89–101. doi: 10.1016/j.cedpsych.2008.09.002

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. *Psychonomic Bulletin & Review*, *18*, 267–277. doi: 10.3758/s13423-010-0042-0

Voyer, D., Postma, A., Brake, B., & Imperato-McGinley, J. (2007). Gender differences in object location memory: A meta-analysis. *Psychonomic Bulletin & Review*, *14*, 23–38.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250-270. doi: 10.1037/0033-2909.117.2.250

Vukovic, R. K., Kieffer, M. J., Bailey, S. P., & Harari, R. R. (2013). Mathematics anxiety in young children: Concurrent and longitudinal associations with mathematical performance. *Contemporary Educational Psychology*, *38*, 1–10. doi: https://doi.org/10.1016/j.cedpsych.2012.09.001

Wai, J., Cacchio, M., Putallaz, M., & Makel, M. C. (2010). Sex differences in the right tail of cognitive abilities: A 30-year examination. *Intelligence*, *38*, 412–423. doi: 10.1016/j.intell.2010.04.006

Wang, M. T., & Degol, J. L. (2016). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*, 119-140. doi: 10.1007/s10648-015-9355-x

Wholey, J., Hatry, H., & Newcomer, K. (2010). *Handbook of practical program evaluation*. San Francisco, CA: Jossey-Bass.

Williams, J. A. (2010). Learning differences: Sex-role stereotyping in single-sex public education. *Harvard Journal of Law & Gender*, *33*, 555–579.

Appendix A

Needs Assessment Survey of Student Math Attitudes

Directions: The following survey contains a number of statements with which some people agree and others disagree. Please rate how much you personally agree or disagree with these statements.

Likert Scale: Strongly agree=4 , agree=3, disagree=2, strongly disagree=1

1. Mathematics is a very worthwhile subject
2. Mathematics is important in everyday life
3. Mathematics is one of the most important subjects for people to study
4. I don't really use math outside of school
5. My mind goes blank and I am unable to think clearly when working with math
6. Studying mathematics makes me feel nervous
7. When I hear the word mathematics, I have a feeling of dislike
8. It makes me nervous to even think about having to do a mathematics problem
9. I have a lot of self-confidence when it comes to mathematics
10. I expect to do fairly well in any math class I take
11. I feel a sense of insecurity when attempting a mathematics problem
12. I learn mathematics easily
13. I believe studying math helps me with problem solving in other areas
14. A strong math background could help me in my professional life
15. I believe I am good at solving math problems
16. I consider myself a "math person"
17. I believe some people are naturally "math people" and others are not
18. In general, boys are better at math and girls are better at reading
19. I feel nervous doing mathematics problems
20. I am just not good at mathematics
21. I worry that I will get bad grades in math
22. I believe people have a certain amount of intelligence, and they can't really do much to change it
23. I believe people can substantially change how intelligent they are
24. I think people can learn new things, but they can't change their basic intelligence
25. I think people can change even their basic level of intelligence considerably
26. I feel valued and appreciated in math class
27. I feel uncomfortable and out of place in math class
28. I like to solve new problems in mathematics
29. Mathematics is dull and boring

30. My gender is (male/female)
31. My grade level is (fourth-twelfth)

Appendix B

Updated Survey of Student Self-Efficacy and Sense of Belonging

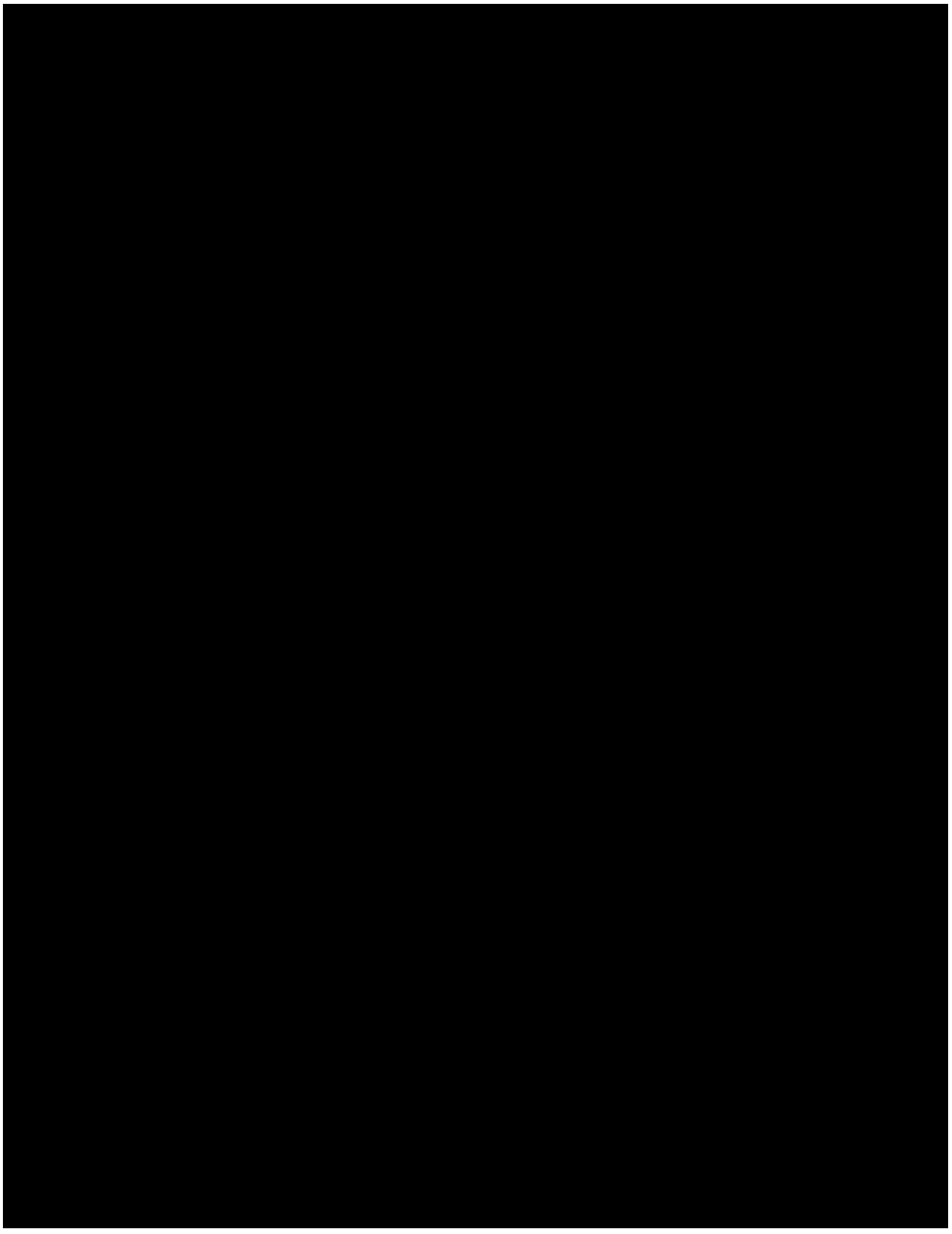
Directions: The following survey contains a number of statements with which some people agree and others disagree. Please rate how much you personally agree or disagree with these statements.

Likert Scale: Strongly agree=4 , agree=3, disagree=2, strongly disagree=1

1. I have a lot of self-confidence when it comes to learning math.
2. I learn math easily.
3. I believe I am good at solving math problems.
4. I feel unsure of myself when trying to solve a new math problem.
5. I'm not the type to do well in math class.
6. Math seems unusually hard for me.
7. I feel valued in my math class.
8. I feel left out in my math class.
9. I feel comfortable in my math class.
10. I feel out of place in my math class.
11. I wish I could fade into the background in my math class.
12. I enjoy being an active participant in my math class.
13. My gender is (male/female/other).
14. My grade level is (5=fifth, 6=sixth, 7=seventh, 8=eighth).
15. My name is _____(text box)_____

Appendix C

RTOP Observation Tool



Appendix D

Pilot Year (2015-2016) Math Workshop Interview Questions

Classroom Culture

		Never Occurred				Very Descriptive
Communicative Indicators	16.) Students were involved in the communication of their ideas to others using a variety of means and media.	0	1	2	3	4
	17.) The teacher's questions triggered divergent modes of thinking.	0	1	2	3	4
	18.) There was a high proportion of student talk and a significant amount of it occurred between and among students.	0	1	2	3	4
	19.) Student questions and comments often determined the focus and direction of classroom discourse.	0	1	2	3	4
	20.) There was a climate of respect for what others had to say.	0	1	2	3	4
Student/ Teacher Relationships	21.) Active participation of students was encouraged and valued.	0	1	2	3	4
	22.) Students were encouraged to generate conjectures, alternative solution strategies, and ways of interpreting evidence.	0	1	2	3	4
	23.) In general the teacher was patient with students.	0	1	2	3	4
	24. The teacher acted as a resource person, working to support and enhance student investigations.	0	1	2	3	4
	25.) The metaphor "teacher as listener" was very characteristic of this classroom.	0	1	2	3	4

Feedback

Appendix E

Pilot Year (2015-2016) Interview Questions

1. Tell me your "math history." What are some of your memories of doing math in school?" Tell me about your experiences doing math at BC

2. What three words pop to mind when you think of learning math?
(why did you choose these words?)

3. Do students in your math classes behave differently from how they behave in other classes? What is the difference? Why do you think that happens?

4. Tell me about your experiences in math workshop. How do they compare to your regular math class?

5. Fill in the blank: when I think about going to math workshop, I feel
_____

6. What do you like best about math workshop? What suggestions do you have for how to make math workshop better?

Appendix F

Intervention Year (2016 to 2017) Interview Questions

1. Do you feel confident in your regular math class? Why or why not?

2. Are there any times when you feel nervous in your math class? Can you give me an example?

3. Do you ever feel uncomfortable or out of place in your regular math class? If so, when? What is an example or situation?

4. What helps you feel comfortable in math class?

5. What was different about math workshop compared to your regular math class? What was similar?

Biography

Susanna Lawrence Brock was born in New York City in 1984.

Susanna attended Harvard College and graduated in 2007. She concentrated in History and Science with a certificate in Mind, Brain, and Behavior studies. Following college, she taught math and science at a number of independent schools including Greenwich Academy in Greenwich, CT and The Convent of the Sacred Heart in New York, NY. In 2011, she earned her Master of Science degree at Columbia Teacher's College in Neuroscience and Education. In 2013, Susanna joined the Berkeley Carroll School in Brooklyn, NY as the middle school mathematics specialist. The following year, Susanna began her Ed.D. at Johns Hopkins University. Her dissertation study is based on her work with students at Berkeley Carroll.